ERRATA

p. 145, Equation 5.2: “R^2” for “r”.

p. 146, line 1 of para 2: “this” for “the is”.

p. 180, line 1 of Section 6.3: “*SWSM2009 Clustering Dataset*” for “*SWSM2009 Clustering Dataset* dataset”.

p. 199, line 3 of para 3: “the chosen measure” for “the chosed measure”.

p. 274, line 7 of Section 8.2: “subsequently” for “sunsequently”.

p. 319, Reference 7: “pp. 145–153” for “p.145153”.

ADDENDUM

p. 39, para 3, first sentence: Delete “supervised”.

p. 142, Figure 5.3: Comment: There is no yellow colour in Figure 5.3. The orange pie slice represents the sum of percentages for the green, purple, and light blue categories (in the secondary chart).

p. 145 Add at the end of para 1:
“Strings containing non-ASCII text, such as Japanese *Katakana* characters, are represented in Figure 5.5 as ‘*(non-ASCII source string)*’. Six such occurrences, found along the abscissa, represent six different non-ASCII `source` strings. (Multiple yet independent occurrences of such strings are commonly found throughout the dataset). The distribution of `source` strings in Figure 5.5 will be discussed at the end of this section.”

p. 145, para 2: Delete “where n is the number of data points, X is the original data set and Y representing the trendline’s data points.”; and read:
“where n is the number of data points, X is the original data set, Y is the set containing the trendline’s data points, s_X and s_Y are the sample standard deviations for X and Y respectively, and $\bar{X}$ and $\bar{Y}$ are the sample means for X and Y respectively.”

p. 145, para 3: Delete “power-low model”; and read: “power-law model (Equation 5.1)”.

p. 145, para 3: Delete “with x being the ranked variables from the original data set, and y the ranked variables from the trendline’s data points.”; and read:
“with x being the ranked variables from the original data set, y the ranked variables from the trendline’s data set, and $\bar{x}$ and $\bar{y}$ the sample means for x and y respectively.”

p. 147, Figure 5.6: Comment: the legends are organized in row-major order, corresponding to the device classes in descending frequency. ‘*Others*’ represents the classes falling within the long tail of the distribution (3.20%).

p. 172, list item 5: Delete “combined with a SOM for”.

p. 177, Figure 6.1: Comment: Figure 6.1 is simply to illustrate the discrete values of the *primary usage pattern* feature. The figure clearly indicates how a given feature prominently appears within a particular cluster; e.g. a primary usage pattern of *personal* is exclusively seen in the bottom-left (largest) cluster. Global SOMs for all features are illustrated in Figures 6.2–6.4, grouped according to the categories of *long-*, *medium-* and *short-term topics*, as defined in Section 6.2.1.

p. 177, Figure 6.2: Delete existing caption; and read:
"Figure 6.2: Final maps for two different *long-term topics* in Table 6.1. (left) Global SOM clusters over all features for the topic `Revolverheld`; and (right) Global SOM clusters over all features for the topic `Nizar`.

p. 177, Figure 6.3: Delete existing caption; and read:
"Figure 6.3: Final maps for two different *medium-term* topics in Table 6.1. (left) Global SOM clusters over all features for the topic `H1N1`; and (right) Global SOM clusters over all features for the topic `TwitHit`.

p. 178, Figure 6.4: Delete existing caption; and read:
"Figure 6.4: Final maps for two different *short-term* topics in Table 6.1. (left) Global SOM clusters over all features for the topic `Grey's Anatomy`; and (right) Global SOM clusters over all features for the topic `coffee`.

p. 180, para 2, second sentence: Delete "The remainder of the sample set (in red) comprises of Twitter user accounts involved in coffee-related marketing campaigns and news aggregation; some of the accounts in this cluster have been suspended or banned based on policy violation."; and read:
"As for the second cluster (in red), its pertinent features include: a high incidence of marketing and news aggregator clients (Table 4.9 contains a complete treatise on such clients); users who are organizations (instead of human Twitter users); and users from undetectable geographic locations. With the prominence of said features, this cluster exhibits characteristics of Twitter spammers dealing with coffee-related marketing campaigns. Some of the accounts in this cluster have been suspended or banned based on policy violation, which vindicates my claim."

p. 188, para 1: Delete "or ISODATA".

p. 189, para 4: Delete "Also, k-means is order-independent, as the same clustering of the data irrespective of the order in which the data is presented to the algorithm [Abbas, 2008]".
Comment: the aforementioned claim was wrongly made by Abbas [2008].

p. 190, para 6: Delete "Depending on how the clusters are initialized, k-means can often produce local optima can be achieved [Baçao et al., 2005]."; and read:
"k-means can often produce local optima, depending on how the clusters are initialized [Baçao et al., 2005]."

p. 209, para 3: Delete "degrees of freedom (d.o.f.) = 18 (Equation 7.1)."; and read:
"degrees of freedom (d.o.f.) $= (n-1) = (19-1) = 18$ (with n being the number of values, as per Equation 7.1)."

p. 233, para 3, second sentence: Delete "Several other machine learning methods such as Bayesian networks, and Support Vector Machines could be employed via data mining packages such as Weka [Hall et al., 2009]"; and read:
"Other unsupervised clustering methods – such as *Expectation Maximization* (EM) and *Cobweb* – could be employed via data mining packages such as Weka [Hall et al., 2009]."

p. 239, Figure 7.14: Delete existing caption; and read:
"Figure 7.14: SOM clustering for the 'Paz Sin Fronteras II' simulation data. The detailed cluster maps illustrate features of interest (from left-to-right): Row 1 – message IDs, fixed devices, mobile devices, retweet messages, and reply messages; Row 2 – hashtagged messages, messages containing images (linked to *TwitPic*), messages containing URLs, authors with undetectable gender, and female authors; Row 3 – male authors, authors' geographic latitude, and authors' geographic longitude."

p. 241, Figure 7.15: Delete existing caption; and read:
"Figure 7.15: SOM clustering for the 'AFL Preliminary Final' simulation data. The detailed cluster maps illustrate features of interest (from left-to-right): Row 1

– message IDs, fixed devices, mobile devices, retweet messages, and reply messages; Row 2 – hashtagged messages, messages containing images (linked to *TwitPic*), messages containing URLs, male authors, and female authors; Row 3 – authors with undetectable gender, authors' geographic latitude, and authors' geographic longitude."

p. 246, second-last para, last sentence: Delete "Twitter users who are concerned about how the riots might affect them (or their potential spread) due to their countries' links to the United Kingdom."; and read:
"Twitter users who are concerned about how the riots might affect them (or their potential spread) due to their personal links to the United Kingdom."

p. 258, Figures 7.23 and 7.24: Comment: In some situations, where e.g. a Twitter user is merely sitting at home tweeting about the riot (categorized in Cluster I by the SOM-Ward algorithm), such a user can have features prevalent amongst rioters (Cluster II). Such features include usage of mobile phones, male gender, and geographic proximity to the riot location [Rogers et al., 2011; May, 2011].

p. 247, Figure 7.16: Comment: the colors used in the map are on a continuous logarithmic scale within the range [0 – 47617]. The colours indicated in the legend only represent the lower bound, median, and upper bound.

p. 317–p.334: Referencing.
Comment: Several references used throughout this thesis do not come with page numbers, as the said conference proceedings/journals are distributed in electronic format (e.g. individual PDFs or webpages). These include, but are not limited to: Proc. *GIR'10*, Proc. *CHI2010 Workshop on Microblogging*, Proc. *HICCS-43*, Proc. *COMSNETS'09*, Proc. *MSM '10*, Proc. *DEIM Forum 2010*, and *First Monday.* The following is a list of references where page numbers have now been made available.

p. 319, Reference to "Burns, A. and Eltham, B. [2009]": Add at the end of reference: "pp. 298–310."

p. 320, Reference to "Choudhury, M. D., Sundaram, H., John, A. and Seligmann, D. D. [2008]": Add at the end of reference: "pp. 55–59."

p. 320, Reference to "Claster, W., Dinh, H. and Cooper, M. [2010]": Add at the end of reference: "pp. 158–163."

p. 321, Reference to "Dunlap, J. C. and Lowenthal, P. R. [2009]": Add at the end of reference: "pp. 129–136."

p. 322, Reference to "Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. [2005]": Add at the end of reference: "pp. 78–87."

p. 323, Reference to "Huang, J., Thornton, K. M. and Efthimiadis, E. N. [2010]": Add at the end of reference: "pp. 173–178."

p. 324, Reference to "Jamali, M. and Abolhassani, H. [2007]": Add at the end of reference: "pp. 58–64."

p. 325, Reference to "Kumar, R., Mahdian, M. and McGlohon, M. [2010]": Add at the end of reference: "pp. 553–562."

p. 325, Reference to "Kwak, H., Lee, C., Park, H. and Moon, S. [2010]": Add at the end of reference: "pp. 591–600."

p. 326, Reference to "Mathioudakis, M. and Koudas, N. [2010]": Add at the end of reference: "pp. 1155–1158."

p. 327, Reference to "Merelo-Guervs, J. J., Prieto, B., Prieto, A., Romero, G., Valdivieso, P. C. and Tricas, F. [2004]": Add at the end of reference: "pp. 158–165."

p. 327, Reference to "Neria, Y., Suh, E. and Marshall, R. [2004]": Add at the end of reference: "pp. 201–215."

p. 330, Reference to "Starbird, K., Palen, L., Hughes, A. and Vieweg, S. [2010]": Add at the end of reference: "pp. 241–250."

p. 331, Reference to "Stoica, A., Couronne, T. and Beuscart, J.-S. [2010]": Add at the end of reference: "pp. 154–161."

p. 331, Reference to "Thomas, K., Grier, C., Paxson, V. and Song, D. [2011]": Add at the end of reference: "pp. 243–258."

p. 333, Reference to "Westman, S. and Freund, L. [2010]": Add at the end of reference: "pp. 323–328."

p. 333, Reference to "Yamazaki, Y. and Kumasaka, K. [2010]": Add at the end of reference: "pp. 357–360."

Inferring Social Behavior and Interaction on Twitter by Combining Metadata about Users & Messages

by

Marc Chi-Yan Cheong, BCompSc

Thesis

Submitted by Marc Chi-Yan Cheong

for fulfillment of the Requirements for the Degree of

Doctor of Philosophy (0190)

Supervisor: Dr. Sid Ray

Joint Supervisor: Prof. David Green

Clayton School of Information Technology
Monash University

February, 2013

This thesis is dedicated to my late uncle, ‘Frankie’ Chan Kok-Khuin.

This is for you, Frankie.

Quoth the words of Elvis Presley... “you are always on my mind.”

Contents

List of Tables

List of Figures

Inferring Social Behavior and Interaction on Twitter by Combining Metadata about Users & Messages

Marc Chi-Yan Cheong, BCompSc
[redacted]
Monash University, 2013

Supervisor: Dr. Sid Ray
[redacted]
Joint Supervisor: Prof. David Green
[redacted]

Abstract

Social media — in particular microblogging — is fast becoming important in today's world. A good example is Twitter, which is a rich source of readily-available information by, and about, people. Real-life happenings are constantly reported on Twitter; thus, it functions as a 'mirror' to the real world. These happenings range from the banal (individual thoughts, opinions, and observations), to the dramatic (celebrity announcements, scandals, and Internet memes), to real-world events with serious consequences (riots, coordination during natural disasters, response to terrorism, and political dissent).

Most extant literature treats the message and user domains on Twitter independently of one another. Current research focuses only on a single domain, but rarely on both. Research consists mostly of specialized techniques, such as opinion and sentiment mining, community detection, social network analysis, and trend mining which are merely applied to Twitter data. Rarely are metadata from both the user and message domains analyzed in tandem with each other. My thesis combines metadata from both domains and transforms them into useful inferences for detecting hidden patterns. The basis of my research is the use of metadata from both Twitter users and messages as the raw material, from which we can discover hidden patterns and inferences. Such patterns and inferences, in turn, can be combined with data mining techniques to unearth a wealth of knowledge about Twitter users in particular, and people in general. In this thesis, I investigate two aspects. First, I introduce a new framework for the large-scale gathering and collation of Twitter user and message metadata. Secondly, I introduce and investigate new inference algorithms that combines metadata from both domains, inspired by current literature, which are hitherto absent in research. In doing so, I contributed to the development of novel inference algorithms, and frameworks to harvest raw metadata from Twitter for the provision of ample data for the evaluation of my algorithms.

From the wealth of metadata from the two domains on Twitter, my new algorithms produce three categories of inferences — social demographics, exhibition of online presence by users, and messaging (*tweeting*) behavior of users. Using these new inference algorithms,

I tested my findings on a large-scale real-world dataset, collected from Twitter using data-gathering frameworks I have developed. Consequently I was able to draw conclusions of the current '*state of the Twitterverse*'. Following that, I introduced a novel application of pattern detection and clustering on inferences generated from my algorithms. This is for the detection of latent traits and identification of non-obvious patterns, with respect to the three categories of inferences that are generated from my algorithms.

To conclude my thesis, I showed that my approaches provide useful insights about serious real-world phenomena captured on Twitter — pertaining to environmental activism, terrorism events, and public disorder — all of which are of interest to researchers, governments, and the media alike. Using the approaches proposed throughout my thesis, I was able to discover the behavior of people in the real world, and illustrated how such real-life behavior is translated into expression and social communication in the online realm. The results from these studies covered in my thesis led to a better understanding of who social media consumers are, how they communicate online, and how behavioral patterns from these users 'mirror' the real-world.

Inferring Social Behavior and Interaction on Twitter by Combining Metadata about Users & Messages

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Marc Chi-Yan Cheong
February 19, 2013

Acknowledgments

First and foremost, I'd like to express my gratitude to my supervisors, Dr Siddheswar Ray ("*Baba*") and Professor David Green, for their counsel, guidance, and mentoring throughout my entire journey through this doctorate. I am proud to count them as not only my supervisors, but also as wonderful colleagues, mentors, and most importantly, as my friends.

I'd like to thank and express my love for my parents, Henry Cheong and Rose Cheah, for their unconditional love, support, and understanding. *Ma*, thanks for instilling in me a passion for learning and a voracious appetite for books. *Low Tow*, thanks for buying my first 486DX computer in '95 (which paved the way to this PhD, 18 years later), and sharing with me the anecdotes of life over father–son Tiger Beer sessions. I love you folks.

To my dear colleagues, thank you for being there for me, in one way or another, in my doctoral and teaching career. From the Sunway campus — Anushia, Bhawani, Choon-Ling, Fariza, and Radi — thanks for inspiring me to start my PhD. To my fellow Melburnian colleagues and *postgraduanators*, it has been an amazing journey with you all — Amiza, Alda, Arun, Ben G, Cameron, Çağatay, Chris, Dhananjay, Fabian, my *FIT1031 Family* (Gopal-*ji*, Iris, Malik, Manoj, Nitin, Subrata), Hiran & Upuli, Kerri, Mauro, Owen, Peter M, Shenjun, and '*The Lunch Bunch*' (Masud, Robert, Ron, Thomâs). *Happy Friday!*

I'd like to especially acknowledge Ben Porter, Brendon Taylor, Marsha Minchenko, Minh-Duc Cao, Robyn McNamara, and Sheikh Mohd Rokonuzzaman. From the bottom of my heart, thanks for being there for me, as loyal confidants and great mates. You have, more than once, stood by me whenever I was in most need of support and a helping hand, and got me back on my feet (and served me Coronas & gin-tonics) whenever life handed me lemons.

A heartfelt thanks to the admin staff — Amelia Morrison, Jack Sotiriou, Karen "*Mum*" Fenwick, Karen "*Kâren*" Nisbet, Kerry McManus, Leeanne Evans, and Wendy Tregenza — for being there throughout my research and teaching career. My acknowledgements also to tech staff Paul van Haaster and the late Gary Evans, for supporting me during the times my computer ran amuck. Also, thanks to Prof. Frada Burstein, Dr. Kar-Seng Loke, and Ms. Shiranthi Ponniah who encouraged me to pursue a doctorate; to A/Prof. Vincent C.S. Lee for the valuable research feedback and input; and my research collaborators Dr. Douglas "*Dougy*" Stones and Dr. Torgeir Watne.

My dear friends — Devendran Raghavan, Hui-Kean & Jeffrey & Kevin & Yeng-Chong "*Corporal*" Lee (*all surprisingly not related!*), Frank Tassone, Nick "*Gawney*" Gawne, Robert Greenaway, Harry & Shaun Olikh, Jess Crawshaw, Siang-Yik Kok, Toby Lu — guys, I made it... time for beer!

To my neighbours and exchange-student friends at *Howitt Hall*, class of 2009–2012 — for all the fun times, and for making me feel young again throughout my stay — Dr 'Marky' Marc thanks you! :)

Last but not least, to the students whom I have been privileged to be lecturer, supervisor, and tutor of — for making my early teaching career enjoyable, for the amazing commendations, and for letting me learn together with you all — you have been a lovely audience!

> "*Lend me your ears and I'll sing you a song,*
> *and I'll try not to sing out of key.*"
> — The Beatles, *With A Little Help From My Friends*

Marc Chi-Yan Cheong

Monash University
February 2013

Vita

Publications arising from this thesis include:

Cheong, M. [2009]. What are you Tweeting about?: A survey of Trending Topics within the Twitter community, *Technical Report 2009/251*, Clayton School of Information Technology, Monash University.

Cheong, M. and Lee, V. [2009]. Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base, Proc. CIKM 2009 Co-Located Work-shops: SWSM 2009, pp. 1–8.

Cheong, M. and Lee, V. [2010a]. Dissecting Twitter: A Review on Current Microblogging Research and Lessons from Related Fields, *From Sociology to Computing in Social Networks: Theory, Foundations and Applications, Vol. 1 of Lecture Notes in Social Networks*, Springer-Verlag, pp. 343–362.

Cheong, M. and Lee, V. [2010b]. A Study on Detecting Patterns in Twitter Intra-topic User and Message Clustering, *Proc. ICPR 2010*, pp. 3125–3128.

Cheong, M. and Lee, V. [2010c]. Twitmographics: Learning the Emergent Properties of the Twitter Community, *From Sociology to Computing in Social Networks: Theory, Foundations and Applications, Vol. 1 of Lecture Notes in Social Networks*, Springer-Verlag, pp. 323–342.

Cheong, M. and Lee, V. [2010d]. Twittering for Earth: A Study on the Impact of Microblogging Activism on Earth Hour 2009 in Australia, *Proc. ACIIDS 2010.*

Cheong, M. and Lee, V. [2011]. A Microblogging-based Approach to Terrorism Informatics: Exploration and Chronicling Civilian Sentiment and Response to Terrorism Events via Twitter, Information Systems Frontiers **13**(1): 45–59.

Cheong, M. and Ray, S. [2011]. A Literature Review of Recent Microblogging Developments, *Technical Report 2011/263*, Clayton School of Information Technology, Monash University.

Cheong, M., Ray, S. and Green, D. [2012a]. Interpreting the 2011 London Riots from Twitter Metadata, *Proc. SoCPAR 2012.*

Cheong, M., Ray, S. and Green, D. [2012b]. Large-scale Socio-demographic Pattern Discovery on Microblog Metadata, *Proc. SoCPAR 2012.*

Permanent Address: Clayton School of Information Technology
Monash University
Australia

This thesis was typeset with $\LaTeX\,2_{\varepsilon}$[1] by the author.

[1] $\LaTeX\,2_{\varepsilon}$ is an extension of $\LaTeX$. $\LaTeX$ is a collection of macros for $\TeX$. $\TeX$ is a trademark of the American Mathematical Society. The macros used in formatting this thesis were written by Glenn Maughan and modified by Dean Thompson and David Squire and Marc Cheong of Monash University.

Chapter 1

Introduction

“*Hello... is it me you’re looking for?*”

— Lionel Richie,
Hello (1984).

Social media play an important role in modern life as they make information accessible to the wider public. Unlike traditional media, social media consist entirely of user-generated content, making the user a central part of the picture; blurring the distinction between *producer* and *consumer*. Anyone who has access to a blog, a YouTube video, a Facebook profile, a Twitter account, or even a resumé on LinkedIn — is a user of social media.

In the topic of social media, microblogging warrants a particular mention. Microblogging might be described aptly as a hybrid of *social networking* and *small-scale blog message publishing*. It allows people to express themselves and communicate with others via the Internet. Microblogging is relatively recent form of social media, having been popularized circa 2006 with the introduction of Twitter.

Twitter is a rich source of readily-available information by, and about, people. Real-world happenings — from one’s idle musings or celebrity gossip, to serious events such as rioting, terrorism and political dissent — are constantly reported on Twitter. In this regard Twitter is a ‘mirror’ of sorts to the real world. Twitter is used by regular people from all walks of life, celebrities, newsmakers, organizations, and ‘traditional’ media agencies. There are even Twitter accounts driven by automated software or ‘bots’ programmed to disseminate spam.

Due to the availability of a wealth of data about Twitter *users* and *messages*, there is growing interest within academia, government, industry, and the media in studying such data to gain a better understanding of everyday happenings. Examples include interpersonal communication patterns, to effectiveness of marketing campaigns, to public reactions toward governmental policies, and the banality of celebrity feuds. Though the relevance of such studies is readily obvious — especially in the humanities and sciences — there is a lack of emphasis placed on studying Twitter holistically as a whole ‘ecosystem’.

To elaborate, there are two distinct parts (or domains) that constitute the Twitter microblogging service: the *user domain* which focuses on Twitter users and their real-world properties; and the *message domain* which deals with the user-generated messages, or *tweets*, traversing the Twitter network. Extant studies on Twitter treat the two domains independently of one another; current research focuses exclusively on either domain, but rarely both. In other words, the bulk of current research consists mainly of specialized studies from either of the two domains. For the message domain, research encompasses opinion and sentiment mining, trend mining, search and information retrieval, among others. The user domain contains research foci such as community detection, social graph analysis, and network visualization.

However, the above specialized studies miss a lot of crucial information due to their limited scopes. Thus, there is a need for studies that discover knowledge by combining both the user and message domains of Twitter. There is a current lack of understanding about how data from the two domains can be combined and harvested, what knowledge can be acquired from such combined data, and how such knowledge can be used and combined to fathom real-world phenomena. Recognizing this need, this thesis aims to solve the following main question: ***given the richness of data in both user and message domains on Twitter, how can we discover new knowledge from a combination of these domains, that would ultimately lead to a better understanding of real-world behavior and events?***

In this thesis, I have addressed several subgoals, that attempt to answer the main question above. These subgoals are:

1. ***Survey the extent of current research on Twitter.*** Before work can commence on answering the main question, I review academic and popular studies currently available with respect to Twitter and related social media technologies. I classify these works into appropriate categories, and highlight gaps in existing knowledge about Twitter. The rationale behind this subgoal is to examine the current body of knowledge for trends in Twitter research, significant discoveries, and any areas needing improvement. In doing so, my thesis can build up on the strengths of existing work while at the same time address issues in current research which are hitherto unanswered.

2. ***Investigate approaches to extract raw data from both Twitter domains as a basis for generating useful inferences.*** This involves first determining the kinds of raw data available from the user and message domains in Twitter, and evaluating their usefulness. I will then investigate potential methods for obtaining such raw data from Twitter, and any potential issues that might arise from the process. With the raw metadata from both users and messages, I will present several novel algorithms and new metrics that produce inferences on real-life demographic properties, online presence, and communication patterns from said users. The purpose behind this goal is to find ways and means of distilling new information from the deluge of raw Twitter data.

3. ***Perform a large-scale real-world study on current Twitter metadata.*** To accomplish this subgoal, I devise frameworks to simplify the collection of real Twitter data using the different interfaces available on Twitter, based on the knowledge attained from **Subgoal 2**. In particular, these frameworks automate the data collection and archiving process, before piping the output to my earlier-proposed inference algorithms. I tested and validated these frameworks by using them to collect millions of records of live Twitter data. Using my algorithms and metrics, useful inferences on the collected data – from Twitter users' demography to their usage habits – are generated and visualized. This subgoal aims to confirm the robustness of my data-collection strategy and inference algorithms when faced with large input data, and also to provide a large-scale observation of the current state of the *Twitterverse.*

4. ***Clustering and pattern detection to reveal hidden patterns and commonalities amongst generated inferences.*** Within the set of inferences generated from metadata, it is possible that a combination of these inferences would lead to a better understanding of the properties of the users and messages examined. This is more advantageous to studying the various inferences individually, viz. demographic properties, online presence, and communication patterns. To reveal latent commonalities and patterns, I apply clustering methods on the inferences resulting from **Subgoals 2** and **3**. I evaluate the performance and suitability of two popular clustering methods — k-means and the Kohonen self-organizing map — on generated Twitter inferences. The underlying rationale behind this subgoal is to discover how the knowledge discovered using inference algorithms and metrics can be clustered to reveal non-obvious patterns that can lead us to a better understanding of the world as seen from the 'lens' of Twitter.

5. ***Perform case studies on the manifestation of real-world events on Twitter — in the form of user and message activity — in order to understand their real-world nature.*** To accomplish this subgoal, I performed studies of two real-world events — online environmental activism (the Earth Hour Campaign, 2009–2012) and mass rioting (London Riots of 2011) — manifested as online Twitter activity. I also introduced a theoretical framework for terrorism informatics powered by user-generated data on Twitter. These two studies involve the curation of raw data found on Twitter pertaining to the events mentioned, the subsequent derivation of inferences, and analyses linking such inferences to real-world manifestations from the events. The latter study focuses on potentially using the outcomes of **Subgoal 2** and **Subgoal 4** for terrorism informatics. In essence, the main objective of this subgoal is to demonstrate how well the approaches proposed throughout the thesis work, in studies involving significant real-world happenings.

Figure 1.1 provides a big-picture overview of the relationships between the five subgoals in ultimately answering my thesis's main question.

To answer the above five subgoals, this thesis is logically structured into several main parts. I shall start with a brief primer of Twitter and its key concepts (Chapter 2). This is

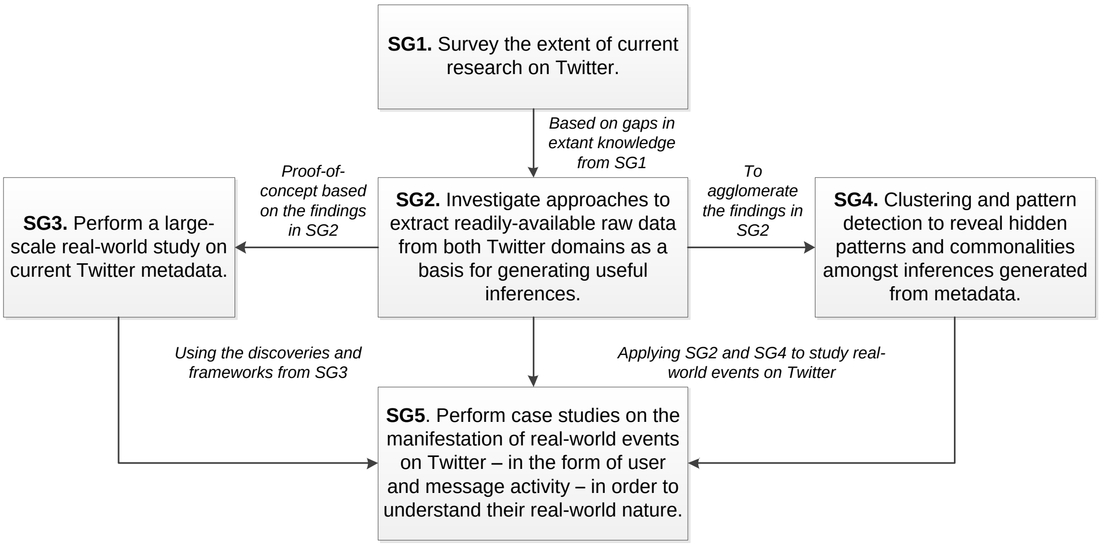

SG1. Survey the extent of current research on Twitter.
Based on gaps in extant knowledge from SG1
SG2. Investigate approaches to extract readily-available raw data from both Twitter domains as a basis for generating useful inferences.
Proof-of-concept based on the findings in SG2
SG3. Perform a large-scale real-world study on current Twitter metadata.
To agglomerate the findings in SG2
SG4. Clustering and pattern detection to reveal hidden patterns and commonalities amongst inferences generated from metadata.
Using the discoveries and frameworks from SG3
Applying SG2 and SG4 to study real-world events on Twitter
SG5. Perform case studies on the manifestation of real-world events on Twitter – in the form of user and message activity – in order to understand their real-world nature.

Figure 1.1: Big-picture overview of the five subgoals in this thesis, their relations, and the logical organization of this thesis.

followed by addressing **Subgoal 1** in Chapter 3, where I have conducted a comprehensive literature review of the state-of-the-art Twitter research. Next, I address **Subgoal 2** in Chapter 4, where I will introduce the interfaces that Twitter provides for extraction of metadata; and the variety of user and message metadata available. I will then detail my novel inference algorithms and metrics which will reveal properties of demography, online presence and communication patterns within said metadata.

Subsequently, **Subgoal 3** is where I introduce two frameworks that automate the process of Twitter metadata collection, post-processing, and storage. This subgoal is addressed in Chapter 5; my frameworks and methods are then put to the test on a large real-world collection of Twitter metadata, which is then used to study the current world-view of Twitter users.

For **Subgoal 4**, in order to reveal latent traits of users and their habits, I investigate the applications of clustering methods on metadata-based inferences (Chapter 6). I also evaluate the suitability of the k-means and self-organizing map approaches when it comes to clustering Twitter inferences.

Chapter 7 addresses **Subgoal 5**: the discoveries from prior chapters are tested on significant real-world events as chronicled on Twitter, where my approaches to knowledge discovery on Twitter are used in tandem with real-world variables and properties that characterize said events.

Having investigated the five subgoals as hitherto listed, in Chapter 8, I document miscellaneous approaches I have contributed to the study of everyday events. These eclectic approaches were secondary discoveries from the point of view of Twitter that have resulted from this PhD.

Finally, I conclude this thesis by reevaluating and reexamining how my subgoals have incrementally led to the answering of my main research question — how knowledge from a combination of both user and message domains lead to a better understanding of real-world behavior and events — as well as presenting future directions of research.

Chapter 2

Background and Key Concepts

> "*I just wanna tell you how I'm feeling,*
> *Gotta make you understand...*"
>
> — Rick Astley,
> *Never Gonna Give You Up* (1986).

2.1 What is Twitter and Microblogging?

In early 2006, Twitter introduced a new form of small-scale web-logging (commonly known as *blogging*), albeit on a much smaller scale [Sarno, 2009]. Hence, the concept of *microblogging* was born: the composition and sharing of short 140-character-limited messages by users. Microblog users can opt to *follow* or subscribe to another user's updates, and vice versa. Each microblog user can update their microblog *profile* (in the case of Twitter, a profile consists of entries such as webpage URL, current location, and profile picture). To summarize, a microblog comprises two core concepts:

1. publishing short messages; and
2. subscribing to (or *following*) other users' messages.

2.2 Microblogging and Twitter: A Brief History

Microblogging has roots in earlier communications systems, such as Internet Relay Chat, from which conventions such as the use of the at-sign (@) and hashtag indicator (#) originated Makice [2009b].

'Away messages,' 'status messages' [Makice, 2009b], or messages that describe the current whereabouts or activities of a user, have become popular after their introduction in instant messaging clients (e.g. AOL Instant Messenger or AIM), and online social networks (e.g. Facebook). These are rather similar, to the current form of microblogging exhibited in Twitter [Levinson, 2009]. Compared to Twitter however, instant messaging is built on the concept of private, interpersonal discussion [Ehrlich and Shami, 2010; Xu and Farkas, 2008] with practically unlimited message length restrictions [Jennings et al., 2006]. Online

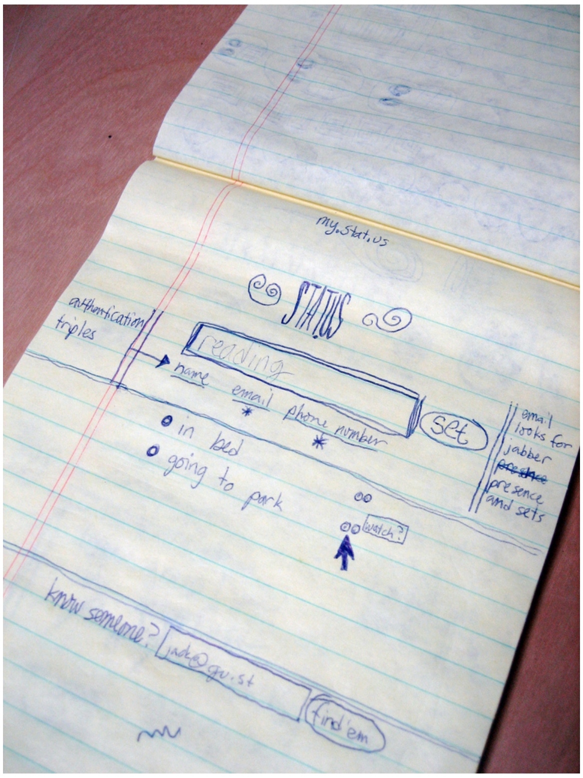


Figure 2.1: A design sketch illustrating *Stat.us*, which was the precursor to Twitter. *Image courtesy of Jack Dorsey, from his Flickr page:* <`http://www.flickr.com/photos/jackdorsey/182613360/`>

social networks on the other hand are based on the idea of building connections with other users, and interacting by means of e.g. applications, profile sharing, tagging, and 'wall posts' [Krishnamurthy, 2009].

Jack Dorsey, Evan Williams and Biz Stone designed Twitter as an in-house communication tool while working in Odeo [Makice, 2009b]. Their team came up with the name Twitter based on their new product's concept: "the physical sensation that you're buzzing your friend's pocket". 'Twitter' was hence coined to replace the original codename of 'status' [Sarno, 2009]. The final product name, according to Dorsey, was "...[defined as] a short burst of inconsequential information, and chirps from birds... that's exactly what the product was" [Sarno, 2009]. Figure 2.1 illustrates an original design sketch by Dorsey, Williams, and Stone for what was to become Twitter.

According to Dorsey, the design for Twitter is inspired in part by the use of the short messaging service (SMS), in particular the 140-character limitation. This character limitation was unique, as explained by Dorsey:

> *...in order to minimize the hassle and thinking around receiving a message, we wanted to make sure that we were not splitting any messages. So we took 20 characters for the user name, and left 140 for the content. That's where it all came from"* [Sarno, 2009].

This, in effect, allows anyone with a mobile phone to use the service — publishing messages, interacting with other users — by way of conventional SMS texts to a Twitter-owned phone number. The concept of *following* is included in the design of Twitter, in the words of Dorsey:

> *...on Twitter, you're not watching the person, you're watching what they produce. It's not a social network, so there's no real social pressure inherent in having to call them a "friend" or having to call them a relative, because you're not dealing with them personally, you're dealing with what they've put out there* [Sarno, 2009].

From its humble beginnings as a novelty status-updating service with the tag-line "*What are you doing?*", it has evolved into a system which is more focused on 'mirroring' happenings in the real world, as illustrated with its updated tag-line "*What's happening?*". Twitter has since grown in terms of its user base, traffic, and the amount of messages it handles. As of the end of 2009, Twitter had approximately 50 million users, with a monthly growth rate of about 16% percent [Moore, 2009; Zarrella, 2009]. In 2012 (as of time of writing), the Twitter user base was estimated at 140 million user accounts [Twitter Inc., 2012b], tripling the previous estimate in a short span of three years.

Twitter's popularity means that a high volume of messages are produced, due to its large user base and ease of use. In 2009, it was estimated at 110 messages per second [Cheong and Lee, 2009]; but during significant world events, the volume of message traffic spikes significantly (increasing by an order of magnitude), evident in the case of the 2011 Women's World Cup Final where a throughput of 7,196 tweets per second was recorded [The Associated Press, 2011]. The high volume of user-generated data flowing through the Twitter service makes it a rich and diverse source of data for research studies.

2.3 Definitions

Several key terms and phrases that explain the core concepts in microblogging have emerged from the lexicon of Twitter users. The following list defines commonly used terms in the Twitter vernacular:

- *Tweet* refers to an individual Twitter microblog message, or the act of composing such a message. The terms *tweet* and *message* are used interchangeably in this thesis.

- *Tweeter* or *Twitterer* are colloquialisms for a Twitter user. To avoid confusion, the term *user* is employed throughout this thesis.

- *Twitterverse* or *Twittersphere* is the user base or community of Twitter users, similar to the term *blogosphere* for the blogging community.

- *Trends* or *Trending Topics* are the top keywords describing the most-discussed topics at a given point in time, automatically ranked by a proprietary algorithm by Twitter.

- *Retweet* or `RT` refers to the forwarding of messages on Twitter, analogous to forwarding an email.
- *Hashtags* (stylized as `#hashtags` in this thesis) are keywords prefixed with a hash sign (`#`), to denote the tagging of a message with said keywords. The generic term *hashtag* and its stylized form are used interchangeably in this thesis.
- `@user` notation addresses another user in a Twitter message; this is typically found in replies, where the user name prefixed with an at (`@`) sign denotes the target recipient of the intended tweet.
- *Following* refers to the act of subscribing to other users' tweets, so that any new tweets posted by the targeted user (*followee*, or Twitter's definition of *friend*) can be visible to the original user who requested (*follower*). For clarity, throughout this thesis, I will italicize the verb "*follow*" and its derivatives ("*follows*" and "*following*") when they are used in the context of 'subscribing to other users' tweets'.

2.4 Twitter Usage, Adoption and Popularity

Twitter has been given wide coverage by the press and has vaulted in popularity to a household topic in recent years. This is due to several events that catapulted it into common knowledge, as well as an increase in adoption of Twitter among the general public. The following briefly describes how Twitter users take advantage of the microblogging service.

2.4.1 Interpersonal Communication

Twitter's original purpose was a platform to inform others, via status updates, about the minutiae of their daily lives [Levinson, 2009]; doing so creates what some call an "ambient intimacy" [O'Reilly and Milstein, 2009] that promotes closer ties with friends and family. This intimacy promotes threaded discussion and chit-chat amongst replying users [Java et al., 2009; Ritter et al., 2010].

2.4.2 Fact-finding and Information Sharing

In addition to personal communication, the vast corpus of public tweets can be 'mined' for information. Users can find solutions to questions, or to get information and news based on the "wisdom of the crowd" [Surowiecki, 2005], or to perform "social search" amongst users in the same social group [Golovchinsky and Efron, 2010].

Concrete examples include using Twitter's crowd wisdom to search for sports scores [Cheong, 2009; Bloch and Carter, 2009], seeking solutions to questions (e.g. job openings) among fellow users [Boyd et al., 2010; Wilson, 2008], getting feedback from people [Naaman et al., 2010] regarding an idea, and searching for locations from peers in a particular geographic location [Kaufman and Chen, 2010]. On the other hand, Twitter users also use Twitter to share and disseminate information; e.g. forwarding links, and retweeting interesting tweets from other users [Boyd et al., 2010; Honeycutt and Herring, 2009].

2.4.3 Public Outreach and Engagement

Twitter has been in the spotlight in recent years due to its role in several-high profile events; most notably its use by then-US presidential candidate, Barack Obama[1], in two successive election campaigns [Harris, 2008]. President Obama has since adopted Twitter in conjunction with other online social networks to reach out to the electorate and provide a platform to connect with young voters [Harris, 2008].

Twitter uses is not only limited to politicians and leaders, but also by celebrities and artists to promote themselves to the world, as exhibited in many popular trends [Cheong and Lee, 2009; Relax News, 2009; Kwak et al., 2010]. The reason is that Twitter's use is common among a young demographic — the current generation of teenagers and those in their early twenties — which forms the target audience for such celebrities. The interaction and social habits of the aforementioned demographic can thus be gleamed by examining their communications on Twitter.

2.4.4 Platform for Collective Action

Politically, Twitter is also used as a means of gauging public opinion and voter preference [Shamma et al., 2009], and as a catalyst for collective action in political activism[Jungherr, 2009; Goolsby, 2009]. Notable scenarios include the 2009 Moldovan 'Twitter Revolution' where Twitter was used to plan a revolution by the people [Serbanuta et al., 2010; Mungiu-Pippidi and Munteanu, 2009]. Twitter also played a role in the 2009 Iranian Presidential Election, where citizens used Twitter to bypass official censorship to express their outrage and to increase global awareness of their cause [Burns and Eltham, 2009; Fleishman, 2009]. In the case of a student activist jailed for expressing dissent, Twitter was a crucial tool to secure his freedom [Simon, 2008]. Twitter has been suggested as a crucial factor for coordinating civil disorder in the 2011 London Riots [Cheong et al., 2012a].

Twitter is also used to break news of crisis and emergency, typically faster than traditional media outlets. Examples of this include the 2009 Hudson River plane crash [Terdiman, 2009], the 2008 Mumbai terror attacks [Beaumont, 2008], and the 2009 Jakarta bombings [Cashmore, 2009a].

Natural disasters such as the 2009 Marseilles fires and the 2009 Red River floods in Canada have seen Twitter activity increase among people affected by the disasters [Longueville et al., 2009; Hughes and Palen, 2009]. The use of Twitter in breaking news of the 2008 earthquake in Sichuan, China, is said to have "started major media coverage and activated responses from world organizations" [Levinson, 2009]. Tweets on suspected earthquakes has been identified as a powerful tool in earthquake epicenter prediction [Guy et al., 2010], and can be used to complement traditional earthquake detection methods.

2.4.5 Leisure

Twitter is also used to spread Internet memes [Dawkins, 1989], or viral information, among the user base. For example, the Twitter community's `#followfriday` meme allows users

[1] President Obama's official Twitter account is located at <`http://twitter.com/BarackObama`>, as of time of writing.

to tweet about other users who are worth *following* on Twitter every Friday, a tradition since the early days of Twitter [O'Reilly and Milstein, 2009]. This allows for the study of the viral spread of information, memes, and online phenomena; complementing existing online studies on the World Wide Web and email [Arbesman, 2004; Wasik, 2009].

2.4.6 Business and Organizations

Twitter is increasingly used by businesses, organizations, and also governments for purposes of public relations, i.e. to promote its accessibility to the public as well as to establish rapport. Politicians use Twitter in promoting rapport with youth [Blossom, 2009], while businesses use Twitter to improve customer relations [Comm, 2009]. Government agencies (e.g. NASA) have also joined the bandwagon, organizing campaigns on Twitter to promote public awareness of their activities [Blossom, 2009; Vertesi, 2010].

People use Twitter as a platform for recommending preferred brands or services [Blossom, 2009; Jansen et al., 2009b]. This point highlights the increasing advantages for businesses to utilize Twitter for acquisition of market intelligence. Twitter is described as a 'listening post' that reveals client sentiment [Comm, 2009].

2.5 User Participation and Privacy

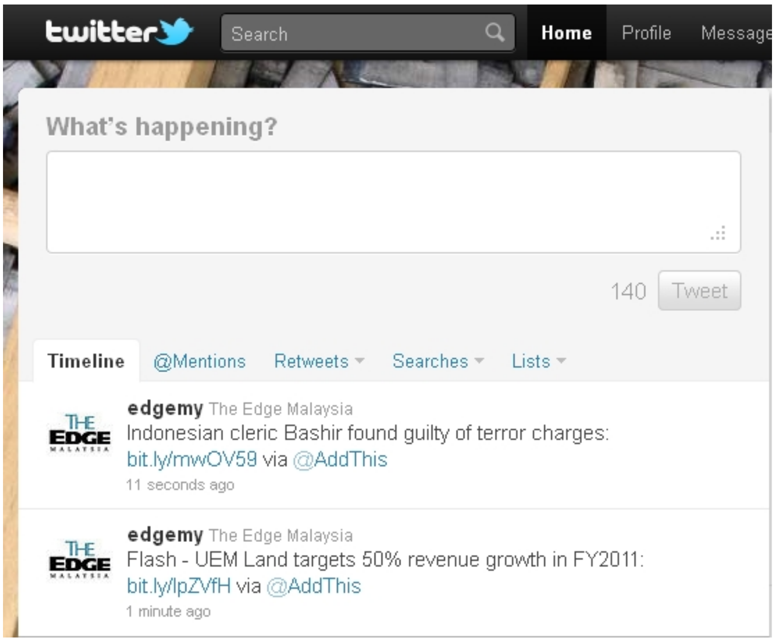
An individual's participation on Twitter is entirely opt-in — a user chooses whether to reveal their presence to the outside world, and whether their tweets are publicly shared with the Twitterverse. Having said that, a significant percentage of the Twitter user base chose to make their Twitter participation public, There are however several documented issues, past and present, with respect to user privacy on Twitter; some of which will be documented later in Section 3.5.3.

Throughout this thesis, I study only public Twitter users who chose to share their presence and tweets to the world. Twitter bars public access to users with private accounts — who are ignored in this thesis — unless such users have explicitly granted permission. From this point onward any description of the Twitter user base in this thesis is restricted to public users.

Twitter is typically used via its web interface, accessible at <`http://www.twitter.com`> (Figure 2.2). An alternative way to use Twitter is Twitter's official mobile service, where users perform tasks by sending short instructions via SMS, one of the main characteristic features of Twitter since its development began [Sarno, 2009]. Recently, there has been a mushrooming of new methods for accessing Twitter [Krishnamurthy et al., 2008; Cheong and Lee, 2010c; Cheong et al., 2012b], e.g.:

Mobile client software These clients are available on modern smartphones (running on e.g. BlackBerry, Android, and Apple iOS platforms). Users directly connect to Twitter via the use of software on these devices (examples are shown in Figure 2.3).

Alternative interfaces These come in the form of new Web 2.0-powered websites or desktop client applications, such as *Seesmic Desktop*, which extend or enhance the functionality of the official Twitter Web interface.



edgemy The Edge Malaysia
Indonesian cleric Bashir found guilty of terror charges: bit.ly/mwOV59 via @AddThis
11 seconds ago

edgemy The Edge Malaysia
Flash - UEM Land targets 50% revenue growth in FY2011: bit.ly/lpZVfH via @AddThis
1 minute ago

Figure 2.2: Screen shot of Twitter's Web interface. Note the presence of the input box for the currently logged-in user to post a tweet, after which is a list of all tweets by the *followers* of the user.

Social-media integrators These allow users to simultaneously and seamlessly participate in Twitter in conjunction with multiple social media services such as Facebook and MySpace [Petrovic et al., 2010; Krishnamurthy et al., 2008].

Sharing tools These programs or web services allow the sharing of not only textual information, but also other media such as images and streaming videos, by augmenting tweets with links to rich media. Compared to *alternative interfaces*, *sharing tools* have a raison d'être of using Twitter for the sole purpose of media sharing.

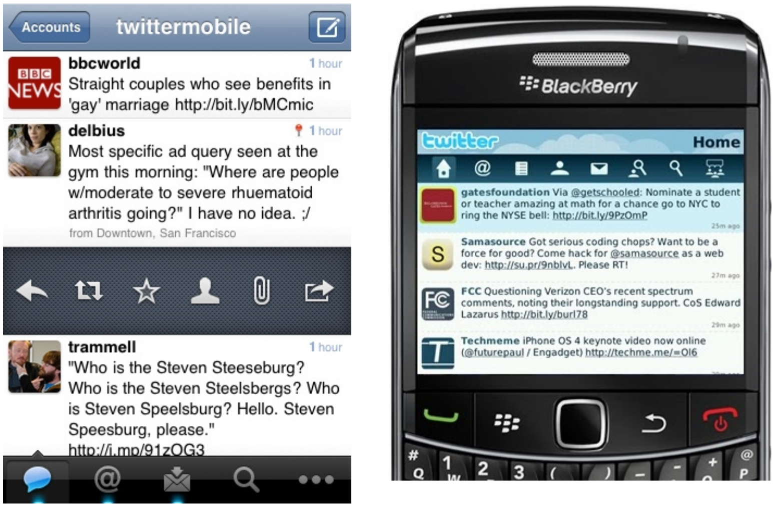
Aggregators These programs read and filter their information from existing online data streams (such as RSS feeds and blogs), and tweet them within Twitter.

Marketing and presence These programs accommodate for Twitter use chiefly in marketing and for 'brand presence', typically by celebrities and organizations. Features include tweeting based on a fixed time interval and the ability to support conccurrent usage by multiple users.

This list is merely a sample of commonly used Twitter access methods or clients. There are many more categories of such clients, which are discussed later in this thesis (Section 4.6.3).

2.6 Central Domains in Twitter as a Microblogging Service

Twitter, when considered as a microblogging service, consists of two interdependent *domains* – the *user* and the *message* – that form the basis of the research presented in this thesis. I introduced these domains in my work in [Cheong and Lee, 2010a], and has since



bbcworld 1 hour
Straight couples who see benefits in 'gay' marriage http://bit.ly/bMCmic

delbius 1 hour
Most specific ad query seen at the gym this morning: "Where are people w/moderate to severe rhuematoid arthritis going?" I have no idea. ;/
from Downtown, San Francisco

trammell 1 hour
"Who is the Steven Steeseburg? Who is the Steven Steelsbergs? Who is Steven Speelsburg? Hello. Steven Speesburg, please."
http://i.mp/91zOG3

gatesfoundation Via @getschooled: Nominate a student or teacher amazing at math for a chance go to NYC to ring the NYSE bell: http://bit.ly/9PzOmP
25m ago

Samasource Got serious coding chops? Want to be a force for good? Come hack for @samasource as a web dev: http://su.pr/9nblvL. Please RT!
27m ago

FCC Questioning Verizon CEO's recent spectrum comments, noting their longstanding support. CoS Edward Lazarus http://bit.ly/burl78
29m ago

Techmeme iPhone OS 4 keynote video now online (@futurepaul / Engadget) http://techme.me/=Ol6

Figure 2.3: Screen shot of Twitter clients for mobile devices: Twitter for the iPhone (left) and BlackBerry (right). *Images courtesy of Apple's Web Store and Twitter Inc.*

been independently formalized as "*central objects*" in a microblogging service [Cormode et al., 2010].

The twin domains of the user and the message are represented internally and conceptually in Twitter as *objects* containing metadata elements (commonly exposed to researchers and programmers in JavaScript Object Notation or JSON[2]). Users and messages are interdependent, and accessible via APIs offered by Twitter to programmers. The two domains are manifested in daily Twitter usage by its end users.

The following subsections briefly describe the two central domains found on Twitter; a concise explanation of the features and metadata found in each domain is discussed in Section 4.3.

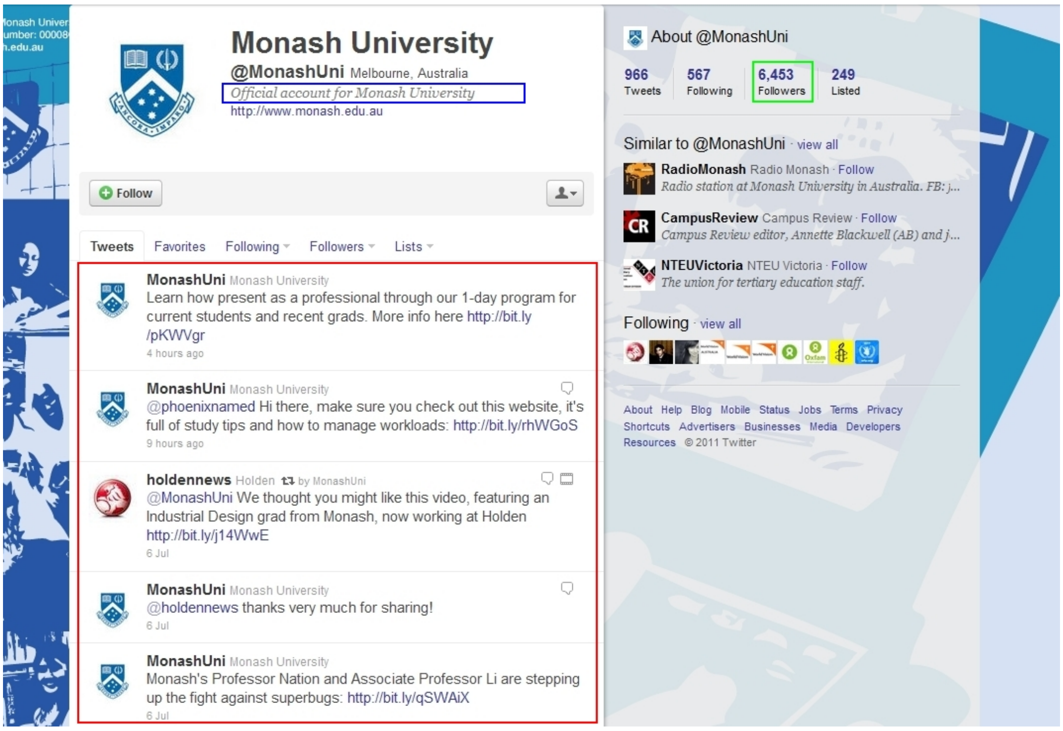
2.6.1 User

The *user* domain models all the characteristics of a user registered on the Twitter service. This domain contains data both directly produced by the user as well as data created as a consequence of user actions. Each user object on Twitter contains metadata comprising basic user profile information (e.g. real name, username, profile description), web profile customization data (e.g. profile picture, profile sidebar color), specific user preferences (e.g. preference of inline image viewing, geotagging), and also statistics calculated by the Twitter service (e.g. verified account status, number of followers, account creation date). Most of these metadata items are directly viewable through the Twitter web interface or using third-party Twitter client software.

More importantly, a given user object links to a collection of messages, i.e. the tweets composed by the user. A user object also has links to other user objects when user-specific actions are performed: for example, when one is *following* another user, or has another user *follow* his/her status updates.

[2] JavaScript Object Notation (JSON) is defined in *RFC 4627* as "a lightweight, text-based, language-independent data interchange format... [which] defines a small set of formatting rules for the portable representation of structured data." <http://tools.ietf.org/html/rfc4627>.

The screen shot in Figure 2.4 illustrates a typical Twitter user, Monash University (`@MonashUni`), as seen through the web interface by another user. Observe how user properties (e.g. description: "`Official account for Monash University`"), summary of *follow*-relationships to other users (e.g. number of followers: "`6,453 followers`"), and a list of all messages (tweets) composed by `@MonashUni` are presented.



Monash University
@MonashUni Melbourne, Australia
Official account for Monash University
http://www.monash.edu.au

MonashUni Monash University
Learn how present as a professional through our 1-day program for current students and recent grads. More info here http://bit.ly/pKWVgr
4 hours ago

MonashUni Monash University
@phoenixnamed Hi there, make sure you check out this website, it's full of study tips and how to manage workloads: http://bit.ly/rhWGoS
9 hours ago

holdennews Holden by MonashUni
@MonashUni We thought you might like this video, featuring an Industrial Design grad from Monash, now working at Holden http://bit.ly/j14WwE
6 Jul

MonashUni Monash University
@holdennews thanks very much for sharing!
6 Jul

MonashUni Monash University
Monash's Professor Nation and Associate Professor Li are stepping up the fight against superbugs: http://bit.ly/qSWAiX
6 Jul

About @MonashUni

966	567	6,453	249
Tweets	Following	Followers	Listed

Similar to @MonashUni · view all

RadioMonash Radio Monash · Follow
Radio station at Monash University in Australia. FB: j...

CampusReview Campus Review · Follow
Campus Review editor, Annette Blackwell (AB) and j...

NTEUVictoria NTEU Victoria · Follow
The union for tertiary education staff.

Following · view all

About Help Blog Mobile Status Jobs Terms Privacy
Shortcuts Advertisers Businesses Media Developers
Resources © 2011 Twitter

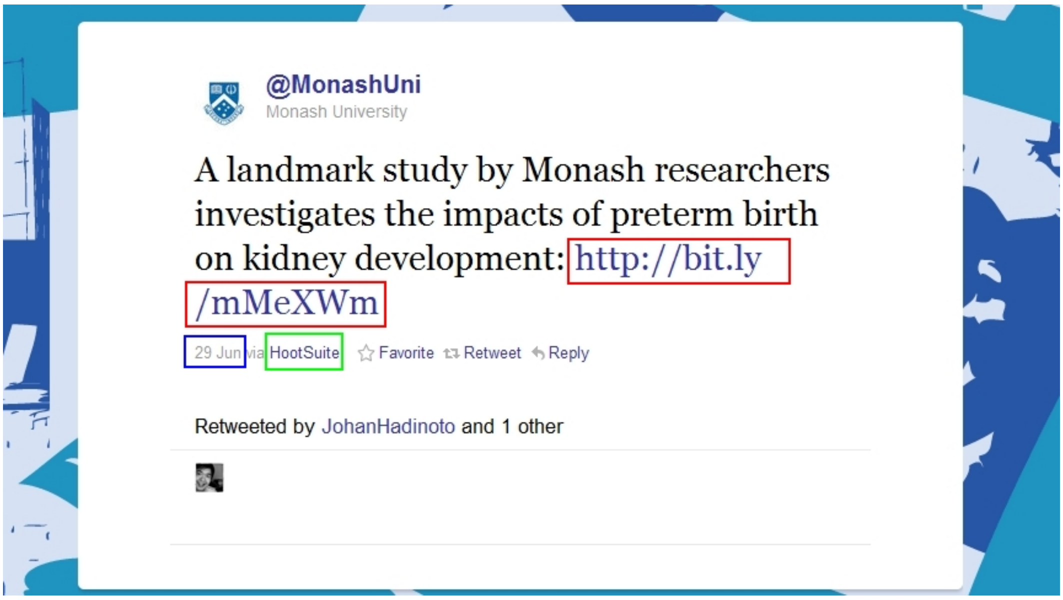
Figure 2.4: Screen shot illustrating the web profile of Twitter user `@MonashUni` (accessible via <`http://twitter.com/MonashUni/`>). Notable features of the profile include user description (blue rectangle), number of followers (green square), and list of tweets (red rectangle).

2.6.2 Message

The *message* domain models a message produced on the Twitter service. A message object belonging to this domain holds the message content itself together with associated data on the composition of the message (such as time created, software client used). Each message object is bound to the creating user object.

Twitter augments the message content with additional information about the message, such as the details of the replied-to message (if the current one was a reply), entities mentioned (users, `#hashtags`, or URLs present in the text), and geographic coordinates of the user at the time a message was sent. All these extra information allows us to determine the context in which a user frames his/her tweet.

The screen shot in Figure 2.5 illustrates a message as seen through the web interface. Key characteristics that can be seen include the author of the message (`@MonashUni`), time stamp ("`29 Jun`"), software client used ("`HootSuite`"), and embedded textual links to additional media or information (<`http://bit.ly/mMeXWm`>).



@MonashUni
Monash University

A landmark study by Monash researchers investigates the impacts of preterm birth on kidney development: http://bit.ly/mMeXWm

29 Jun via HootSuite

Retweeted by JohanHadinoto and 1 other

Figure 2.5: Screen shot illustrating a message by Twitter a user (`@MonashUni`) seen through Twitter's web interface accessible via <`http://twitter.com/MonashUni/status/85925189522694144/`>. Notable features of the message include the time stamp (blue rectangle), software client (green rectangle), and embedded text URL (red rectangles).

2.7 Concluding Notes

This concludes a brief overview of Twitter's history, and also its key concepts, from the point of view of a microblogging service. The primer presented in this chapter provides context for all Twitter-specific discussion throughout the rest of this thesis.

The next chapter (Chapter 3) will house a comprehensive literature review detailing the current research on Twitter and social media in general. Related research — such as human factors in social media, and applications of existing research topics adapted to fit the scope of microblogging — will also be discussed. Within Chapter 3, the state-of-the-art is classified, critiqued, and any gaps or weaknesses found within is pinpointed.

Chapter 3

Current Research

"*Don't know much about history,*
Don't know much biology,
Don't know much about [a] science book,
Don't know much about the French I took..."

— Sam Cooke (as popularized by Herman's Hermits),
Wonderful World (1960).

Parts of this chapter have been published as:

Cheong, M. and Lee, V. [2010a]. Dissecting Twitter: A Review on Current Microblogging Research and Lessons from Related Fields, *From Sociology to Computing in Social Networks: Theory, Foundations and Applications, Vol. 1 of Lecture Notes in Social Networks*, Springer-Verlag, pp. 343–362.

Cheong, M. and Ray, S. [2011]. A Literature Review of Recent Microblogging Developments, *Technical Report 2011/263*, Clayton School of Information Technology, Monash University.

With Chapter 2 covering a bit of the history, evolution, and current popular usage of Twitter as a microblogging service, I now turn my attention to existing research in academia dealing with Twitter and its ilk. This accomplishes **Subgoal 1** of my overall thesis; and sets the stage for the rest of the thesis. My contributions in later chapters will build upon some of these prior research areas, in order to close the gaps and fill the niches identified in the literature.

This chapter details existing academic research done on microblogging in recent years, beginning circa 2007, with special focus on Twitter. Such research is complemented by studies on subjects closely related to microblogging, such as interpersonal communication, the social sciences, human factors, and applications of existing research topics on the two domains of Twitter.

To divide the discussion of the state of the art, I propose a classification methodology, as hitherto detailed in Section 2.6. This classification methodology has been used in my published literature reviews [Cheong and Ray, 2011] and [Cheong and Lee, 2010a] to describe extant research. A piece of research can belong to either (or both) of the two domains found in Twitter and microblogging services:

1. **user domain**: metadata exhibited by a user in a microblogging environment accessible via the Twitter API; this includes statistics such as tweet count, account age, user customization and so forth, which allows the study of human factors behind microblogs.

2. **message domain**: properties exhibited by a single message composed by a Twitter user. The raw data extractable from Twitter API for example includes the message content, software client used, time-stamp, geo-location properties, and any embedded content; which embodies all the characteristics of an individual message.

This chapter explores all the research on and about Twitter, and is organized based on one of the six general themes I have identified in microblogging research:

- Initial exploratory studies on Twitter focus on a thorough exploration of the Twitter service, its features, and idiosyncrasies (Section 3.1).

- With the spread of information on Twitter, some form of self-organization often takes place. This manifests itself especially during significant crisis or large-scale events in the real world: on Twitter, one can observe an epidemic spreading of information due to the actions of its users (Section 3.2).

- From the epidemic spread of information, one is also able to reveal emergent behavior arising from user interactions and/or message activity on Twitter. This is accomplished through the use of pattern detection and clustering methods (Section 3.3).

- The contents of tweets, from the message domain of Twitter, can also be used for modeling, personalized user recommendations, sentiment detection, and user search (Section 3.4).

- Human factors play an important role in the user domain of Twitter, with regards to user intentions and behavior. In particular, parallels exist between online microblogging and human communication in the real world. Hence, I will also review how people tend to communicate and share social information and presence, and the privacy pitfalls involved in such communication (Section 3.5).

- Finally, I look at practical applications of Twitter, particularly in the fields of visualization, computer-human interaction (CHI), and practical applications in and by organizations (Section 3.6).

3.1 Exploratory Studies

The first theme in this literature review is the exploration of the entire Twitter 'ecosystem'. Within this section, I'll provide current facts, figures, and statistics to better understand the current state of Twitter. Also, idiosyncrasies and properties of Twitter messages and users are highlighted, in order to emphasize the uniqueness of Twitter (and Twitter research) compared to other forms of social media and online social networking.

3.1.1 Measurement and Statistics

Several papers [Pear Analytics, 2009; Krishnamurthy, 2009; Cormode et al., 2010] have reported on current measurements and statistics of Twitter in its entirety, from both the user and the message domains. This section discusses these papers and highlights important findings.

Pear Analytics [2009] has published statistical reports on Twitter activity, growth and similar metrics. Vital points in their August 2009 report (the most current, as of time of writing) are summarized in Table 3.1.

Feature	**Statistics**
Users per month (USA)	27 million (June 2009)
Demographics	Gender: 55% female Age group: 43% in the (18-34) age bracket Ethnicity: 78% users of Caucasian descent Annual income: US$30-60k on average
Distribution	Top users: 1% of top users contribute 35% of visits Activity: 72% 'passers-by' as opposed to 27% regular users
Types of tweets	Mainstream news Spam Self-promotion of businesses Babble (everyday trivium): 40.55% of total posts Conversations: 37.55% of total posts Pass-along messages (retweets)
Tweets by time of day	Morning: retweets Midday: news and babble (spam/promotions peaking twice) End of work day: conversations
Tweets by weekday	Early week: retweets Mid-week: news and conversation End of work week: spam and babble

Table 3.1: Pear Analytics [2009]: Vital statistics on Twitter, from August 2009.

From my interpretation, the characteristics summarized above approximate those of a typical middle-class user (cf. Bozkir et al. [2010]). These characteristics describe users in their twenties; whereas younger users in general tend to use other online social networks such as Facebook [Bozkir et al., 2010; Lawler and Molluzzo, 2011]. Another concern found in this report, partially backing this claim, is that Twitter adoption is less frequent among younger people as it is deemed "not safe" due to the lack of "ability to select who they want to connect to" [Pear Analytics, 2009] as opposed to *de facto* social networks such as Facebook and MySpace, which allow the user to control the degree of information shared. This issue of 'selective privacy' will be discussed in greater detail in Section 3.5.

Krishnamurthy [2009], in his discussion of challenges in measuring online social networks (OSNs), characterized Twitter in the Web 2.0 context as a *micro-online social network (micro-OSN)*. Twitter is classified as a micro-OSN because it merely has a limited subset of features of a proper OSN [Krishnamurthy, 2009]. Several features of interest (in parentheses) are detailed in Table 3.2.

Feature class	**Feature**	**Twitter's implementation**
Profile details	*Age/Gender*	None (can be predicted/conjectured in research)
	Location	User text or GPS coordinates
	'Testimonials'	None (closest is a summary in user timeline)
Connectivity	*Friends*	Non-mutual 'followers' relationship
	Subscriptions	Non-mutual 'friends' (*following*) relationship
	Groups	Lists API
Content	*Main*	Microblog entries: 140-character tweets
	Other	Linked URLs in text
	Tagging	Hashtags
	Friends only	None (either completely private or completely public)
	Comments	Tweet reply, using `@user` notation
	Editing	None (can only author/delete)
	Rating	None (closest is implicitly 'favoriting' or retweeting)

Table 3.2: Krishnamurthy [2009]: Twitter features as a Web 2.0 micro-OSN.

Krishnamurthy [2009] also mentioned several unique properties of Twitter that are not obvious in other studies previously covered: presence of cultural bias (as a few countries dominate in terms of geographical Twitter user distribution), of language bias ("Japanese users tweet in Kanji and do not have many English speakers as followers"), and intra-European cliques in terms of user communication patterns.

On a related topic, Cormode et al. [2010] presented a 'manifesto' for modeling and measurement of social media, and discussed several challenges and observations for Twitter as a micro-OSN. Twitter has unique relationships between its *entities* [Cormode et al., 2010]: friends, followers, hashtags, replies, retweets (briefly defined in Section 2.3). This is markedly different from the classical model of networks (comprising nodes and edges), as each type of entity (e.g. *friends*) has different relationships to others, e.g. *retweets*. Also, Cormode et al. [2010] discussed challenges on Twitter research in general, such as the various ways to access data (API versus HTML-scraping versus traffic-sniffing) and sampling methodology (e.g. random node identification versus bounded crawls).

Several issues regarding quality of data have also been identified in the perspective of Twitter: presence of dormant users distorting Twitter friend/follower network properties; constant redesign (such as the newly-introduced *Lists* feature as of end 2009 and the changes in API, which will be discussed later in Sections 4.1 and 4.2); and usage of existing features "...in ways that are unanticipated" [Cormode et al., 2010]. The most striking example of the latter is the "...formalization of previously unsupported conventions adopted organically by Twitter users" [Cormode et al., 2010] such as retweets and hashtags.

Another example is the usage of celebrity pages in place of personal profile URLs. Their case study involving Twitter has formalized the notion of users and messages being "central objects" in Twitter [Cormode et al., 2010], justifying the dichotomy I proposed earlier this chapter (and published in [Cheong and Lee, 2010a]).

As a concluding note, the authors hinted at the limitations of the Twitter API for future measurement and research of Twitter, such as the recently available Twitter stream API making only a sample of the tweets available, the difficulty of measuring the fraction of private tweets, the absence of a dedicated 'grouping' function, and the bias in observations caused by new users and spammers [Cormode et al., 2010].

3.1.2 User Base: Properties, Emergent Features, and Evolution

From facts and figures about Twitter in general, I now evaluate its user base in terms of its key properties, and how it has evolved in recent years. The earliest known study conducted specifically on Twitter and microblogging was by Krishnamurthy et al. [2008]. Krishnamurthy et al. [2008] focused on ",,,distinct classes of users and their behaviors, geographic growth patterns... and [current as of August 2008] size of the network" [Krishnamurthy et al., 2008]. They analyzed the message domain via the public timeline, and crawled the users via out-degree links (i.e. the *follow* relationship), for:

- characterization of users based on their in-degree versus out-degree ratio;
- characterization of users based on the client used to publish tweets (limited as of August 2008) and the timestamps of tweets;
- examination of users' geographic properties via their UTC time zone and domain name; and
- estimation of the size of the entire Twitter user base (at time of writing, August 2008) at approximately ~1.4 million accounts.

In a similar vein, a technical report by Huberman et al. [2008a] has revealed that, as of December 2008:

- **User intention**: Almost 25% of posts are in the form of `@user`, indicating addressivity of messages to friends/followers.
- **Distribution of total tweets**: The number of tweets composed by a user increase to an asymptotic limit as the number of followers (in-degree) increases; the same holds true for the increasing number of friends (out-degree), but reaching no asymptote.
- **Social networking behavior**: Their findings conclude that Twitter users have "two different [social] networks" on Twitter, made up of a dense one of followers/friends, and a sparser one for actual friends. This duality is corroborated by other researchers in more recent work (Section 3.5.3).

Java et al. [2009]'s paper, which chronologically follows [Krishnamurthy et al., 2008] and [Huberman et al., 2008a], and has a broader coverage than the two 2008 studies, contained the following investigations:

- **Growth rate**: Java et al. [2009] estimated user growth to be ~1 million User IDs per month, and message growth rate ~40 million Message UIDs per month (April-May 2007).

- **User distribution**: Twitter has the same power law of degree distributions (in-degree versus out-degree in the user *follow* relationship) as with the Web and conventional blogs.

- **Geographic spread**: Java et al. [2009] determined the various countries of Twitter users based on the usage of GPS coordinates by users who have clients that use GPS coordinates as location information.

- **Community detection and user categorization**: Opinion leaders, general users, and information seekers in each 'community' or subset of the user network were determined based on their in-degree versus out-degree and their message frequency.

- **Frequent trends in a particular community**: Trend keywords are extracted by the authors [Java et al., 2009] by looking at emergent topics in the message domain.

- **User intention**: Analysis was performed on the message content to determine: patterns of chatter, communication via the `@user` notation, information sharing using URLs, and news reporting via RSS feeds.

Chronologically following from Java et al. [2009] is a thesis by Schafer [2010] on the characterization of users on Twitter; and a thorough study of Twitter users and messages by Kwak et al. [2010] to answer the fundamental question of whether Twitter is a "...social network or a news media" [Kwak et al., 2010].

The thesis by Schafer [2010] had a focus using the general characterization of overall Twitter users in order to spot anomalies. Several noteworthy findings, some of which agree with earlier findings from [Java et al., 2009], include:

- **User distribution**: a "clear power law relationship" [Schafer, 2010] was observed in the user graph's degree distribution; corroborating the findings by Java et al. [2009] above.

- **Likelihood of user activity**: approximately 25% of users who registered before 2009 are likely to be active as of 2010; the likelihood that Twitter users who have registered post-2009 still active as of 2010 is only ~15% [Schafer, 2010].

- **Geographic spread**: from the time zone information, Schafer [2010] detected a majority of users from the United States, with significant user bases in Europe and South America. However, developed Eastern nations (Japan and Korea) have a higher number of tweets despite their small user base [Schafer, 2010]. This approximates the detailed findings by [Java et al., 2009] as well.

Kwak et al. [2010], on the other hand, claimed to have surveyed almost the entirety of the Twitter user space, with approximately 41.7 million messages, 1.47 billion friend/follower links, and 106 million tweets. This is achieved via a cluster of 20 machines, each limited to 10k API requests per hour to avoid violating Twitter terms of service (their harvesting includes ten Trending Topics every five minutes; and 1500 tweets are harvested in the same period of time). A summary of their findings is as follows:

- **Spam identification:** The authors have devised a sanitization method to remove likely spammers and 'noise' from their data: users with account ages of less than one day, and tweets which contain three or more Trending Topic mentions are likely to be spam capitalizing on the discussion of a particular topic [Kwak et al., 2010]. In their dataset, 20.2 million such messages and 1.9 million such users have been flagged as such. Also, irregularities in the distribution of number of followers versus number of authored tweets for a given user can be conjectured as being likely to be a spammer's account [Kwak et al., 2010].

- **User network topology:** Based on topological analysis, Kwak et al. [2010] have identified that top users with more than 10k followers are only celebrities or politicians; and those with more than 1 million followers are mainly celebrities and media outlets (such as CNN). The majority of the users with less than 10 followers never contributed a single tweet, while there are also users who "tweet far more than expected from the number of followers" [Kwak et al., 2010], judging by their averages. Another finding is that approximately 67.6% of users are not *reciprocal friends*[1] with a proportion of their followed users — Kwak et al. [2010] suggest that such users are 'consumers' of Twitter as an information source. Finally, Kwak et al. [2010] studied the topology in terms of user homophily and determined that users with 1000 followers or less are geographically close (based on the time zone property of their user accounts) to their reciprocal friends and have a similar measure of popularity (in terms of number of followers, not necessarily reciprocal). Their findings from this part of the survey concluded that Twitter diverges from normal online social network traits as link frequencies are not power-law distributed (due to outliers), users have a short degree of separation[2] on average, and not all friend/follower links between users are reciprocal [Kwak et al., 2010].

- **User rankings:** The authors first applied the *PageRank* algorithm to their user graph to study user ranking from a network perspective, or "propagation of influence" [Kwak et al., 2010]. Next they performed an analysis of user popularity in terms of number of retweets; top retweeted users are politicians and musicians, complementing findings from [Petrovic et al., 2010]. Interestingly, when both sets of

[1] *reciprocal friends*: defined by Kwak et al. [2010] as a Twitter user who is mutually followed by another user he/she is *following*; this term is used throughout Kwak et al. [2010] to define such a bidirectional user relationship on Twitter.

[2] *degree of separation*: defined by Kwak et al. [2010] as the minimum number of connections between a user to any other user; a connection is characterized on Twitter as having a friend and/or follower relationship.

users are compared using a generalized Kendall's tau, they found that there is a discrepancy between the number of followers and the popularity of retweets, bringing "a new perspective in influence" in terms of Twitter [Kwak et al., 2010].

- **Retweeting behavior:** Finally, Kwak et al. [2010] discovered that the "...distribution of the users in a retweet tree [graph representation of retweets] follows [the] power-law distribution" [Kwak et al., 2010] despite the user friend-follow connections not strictly adhering to a power-law. The median of a message retweet is less than an hour, and that retweets are generally diffused rapidly after the second level of retweets (i.e. a retweeted message itself being retweeted) [Kwak et al., 2010]. Fifty percent of retweets happen within the first hour; 75% within a day; and 90% within a month. Favoritism in retweets is evident in that a limited subset of a users' followers actually retweet the message; and that despite the follower count of the original tweet's author, the tweet is "...likely to reach a certain number of audience, once the user's tweet starts [being retweeted]" [Kwak et al., 2010].

3.1.3 Message Domain Properties

Similar to the evaluation of the key properties and evolution of the Twitter user base, I will similarly review the literature for the same regarding the Twitter messages instead. Hence, the following provides facts and figures in extant research that characterize the entirety of the messages found on Twitter.

To start with, I will briefly highlight a study which draws analogs between microblogging and the ancient art of diary-writing from the eighteenth and nineteenth centuries. Humphreys [2010] observed four parallels that exist between current microblogging practices to the diaries of yore:

1. they are both semi-public in nature;
2. they both introspectively chronicle activities and mundane day-to-day trivium;
3. they are both in the narrative form (e.g. entries which discuss upon tragedies); and
4. diary entries are rather short due to limitations of space, similar to microblogs' 140-character constraints

One of the obvious differences is that the diaries lack social interaction; by contrast, thanks to technological advancements, microblog users can now mutually *follow* one another.

Petrovic et al. [2010] have made available the *Edinburgh Twitter Corpus* of approximately ~96.3 million Twitter messages for study and analysis. However, as of time of writing, it has since been removed due to legal issues with Twitter Inc. Nevertheless, Petrovic et al. [2010] managed to summarize the most popular users and most common content on Twitter throughout their survey, which took place in November 2009–February 2010. In their corpus [Petrovic et al., 2010], they have observed the following characteristics:

- Six out of ten of the top followed users were musicians or singers (e.g. `@justinbieber`) building a fan base on Twitter.

- The top hashtagged topics on Twitter dealt with musical-based memes (e.g. `#nowplaying`) where users discuss music they are currently playing; political memes (e.g. `#tcot` or top conservative politicians); tags indicating Facebook co-usage with Twitter; and 'just-for-fun' Internet memes.

- ~80% of the top Twitter client applications are the web interface, *UberTwitter* (indicating a high mobile device usage), and *TweetDeck* (indicating usage of third party applications to improve Twitter user experience (see also Section 4.6.3 later in this thesis).

To conclude this section on message properties and emergent features, I cover a paper by Yoshida et al. [2010], in which the frequency of URLs in tweets and the proliferation of bot-generated content were analyzed. By using data from both the Twitter public timeline and a custom dataset incorporating Japanese links, Yoshida et al. [2010] obtained approximately ~19.7 million tweets containing ~20 million URL mentions. After filtering for non-existent webpages and duplicates, the total number of tweets amounted to ~12.7 million [Yoshida et al., 2010]. By analyzing the source (software client) used to contribute tweets, the authors observed that the web interface, Twitter API-based bespoke programs, and RSS feeds top the list.

Yoshida et al. [2010] also analyzed the tweets for the proportion of software bot-generated tweets, and found a ~35.02% ratio in such automatic postings in their Japanese link dataset (~9.01% for their public timeline samples for comparison). One key distinguishing factor between bot-posted tweets and human-posted ones is that after excluding the URL string in a tweet, bot postings have a higher average string length (48.89 characters) compared to human postings (41.51 characters) Yoshida et al. [2010]. For bots which usually truncate the length of posts, the distribution of tweet lengths have a high peak at the right side of the graph (about 110-120 characters out of the theoretical maximum of 140 *sans* the URL string). By contrast, however, retweets for human posts are higher (~12.55%) compared to bot posts (a mere ~1.44%). Content-wise, the most frequently shared URLs typically consist of photo-sharing websites (the highest URL count, found in ~11.9% of human-posted tweets), media sharing, and news agencies which form the bulk of retweeted content Yoshida et al. [2010] .

3.1.4 Friending, *Following*, and Addressing Behavior

I now shift the focus of the literature review from looking at the 'big picture' of the message domain on the whole, to research on individual users' participation on Twitter. This discussion starts off with academic studies on inter-user activities on Twitter.

Recall in Chapter 2 that Twitter's social aspect lies within a user's ability to *follow* other users. Hence, the focus of this subsection is on the *following* behavior of Twitter users, as well as how they address each other in conversation.

Baumer and Leis [2010] studied the "genres of participation in [Twitter] *following*": they note a shift in media consumption patterns from classical blogs to microblogging; and the existence of an inherent asymmetry of interaction, from the author to the readership of his material. They classified Twitter *following*-patterns into the two categories of *minimalists* and *zealots*:

1. **Minimalists** tend to have 10-30 close friends, use Twitter to socialize, and are normally recommended by others to join Twitter. Such participants "see Twitter as a more intimate... way of connecting" [Baumer and Leis, 2010] with people they are interested in.

2. **Zealots** on the other hand tend to follow hundreds of people, comprising of friends, colleagues, and information sources. To them, Twitter is mainly for "professional and information-seeking purposes" [Baumer and Leis, 2010], and generally adopt Twitter by themselves without being recommended by others. Also, they tend to experiment with Twitter from a variety of software clients.

In their qualitative user study, Baumer and Leis [2010] concluded that despite the differences in user behavior, both zealots and minimalists never use the Twitter Trending Topics feature, and that there seems to be a 'lag time' after signing up to Twitter before their usage became regular [Baumer and Leis, 2010].

A similar study by Heil and Piskorski [2009] illustrated that Twitter users are more likely to follow others from the same gender. Male users are twice likelier to follow other male users; females are 1.25 times more likely to follow other female users [Heil and Piskorski, 2009]. The authors also found that 10% of top Twitter users contribute to about 90% of the content — i.e. the adherence to Zipf's law — which starkly contrasts with a typical online social network where typically, top 10% of users contributing 30% of the content instead [Heil and Piskorski, 2009].

Heil and Piskorski [2009] surmise that Twitter's friend/follow distribution is more of a 'one-way publishing' pattern rather than a social network connecting peers, which is similar to the findings in Kwak et al. [2010]. Studies in political affiliation in the US by Metaxas and Mustafaraj [2010] also revealed that users tend to follow similar users (by political orientation in their study), as the political affiliations of the top 200 users in their dataset is correctly predicted about 98% of the time, simply by observing their *following* habits.

As for addressing and reply-based behavior, indicative of traits of user-directed communication Java et al. [2009], Honeycutt and Herring [2009] performed research into the usage of such `@user` reply messages, and into the existence of coherent conversation patterns among Twitter users. Their study focuses primarily on the message domain, using the direction of messaging to identify threads of conversation using the `@user` messaging pattern.

Honeycutt and Herring [2009] concluded by noting that approximately 91% of messages with a '`@`' sign are intended to signal correspondence between users, and suggests that although Twitter was originally meant to be used to publish status updates, it can indeed be adopted as a platform for purposes of conversation and collaboration. It is important to note that the work by Honeycutt and Herring [2009] is one of the first papers (in early 2009) to study Twitter as a platform for conversation; their findings were vindicated by the high usage of Twitter for interpersonal communication, as seen in Boyd et al. [2010].

3.1.5 Hashtagging and Retweeting

To finish this section on exploratory studies of Twitter, I now present findings conducted with regards to conversational hashtagging and retweeting by individual Twitter users, having read about individual user behavior in the prior subsection.

Huang et al. [2010] performed a study on conversational hashtagging on Twitter. While their study deals with interpretation and statistical trend analysis of hashtags, their analysis of trend from a qualitative aspect will be covered in Section 3.3.2. With respect to trending topics on Twitter (which was not available at the beginning of Twitter's launch), Huang et al. [2010] remarked that "the act of tagging a tweet increased the likelihood of a tweet being... [collated and] displayed in a group of tweets on a trending topic" [Huang et al., 2010].

Hashtagging is a form of emergent behavior by users to tag messages, which developed without any intervention from Twitter staff [Huang et al., 2010; Makice, 2009a]. This is a form of 'conversational tagging', where the tag itself "...is an important part of the message" [Huang et al., 2010] as opposed to merely describing a message. Hashtags also sometimes turn into emergent *micro-memes*, in that users are more inclined to comment or share their views/commentary about a hashtag (topic) only *after* seeing that a particular hashtag has trended.

Boyd et al. [2010] expanded upon Honeycutt and Herring [2009] by studying the 'conversational aspect' of retweeting in the message domain, detailing the forwarding of messages. The stated goals of the paper are to "describe and map out [retweeting] conventions [and] examine retweeting practices" and also draw a similarity between "link-based [conventional] blogging" [Boyd et al., 2010]. A summary of their findings follows.

- **Convention**: Retweeting has no common convention (as of 2009)[3].
- **Content and cascading retweets**: Retweeted messages have elements of information sharing and social tagging, as in the presence of URLs and hashtags. Cascading retweets, akin to cascading forwards in email are also common.
- **Retweet motivations**: The main motivations of retweeting include spreading tweets, to start a conversation, and to draw attention to the originating user. Collective action (e.g. to promote awareness or crowd-sourcing to find answers to problems) are also a motivation for users to retweet messages [Boyd et al., 2010].

[3] A new *Retweeting* feature in the Twitter API was only introduced to Twitter towards the end of 2009. This is of interest primarily to Twitter API developers and researchers, as its introduction was to simplify the notion of user retweets. Further discussion about this feature takes place in Section 4.1.4, later in this thesis

The work by Boyd et al. [2010] adds to the findings of Honeycutt and Herring [2009] by incorporating elements of message retweeting in conjunction with studies on message addressivity and conversation.

3.2 Information Spread and Self-Organization

The second theme of research to be covered in this literature review chapter is how information spreads across Twitter users via tweets as its conduit. The areas that will be covered with respect to this theme include how information spreads in *crisis* and *convergence* [Hughes and Palen, 2009] events: in essence how users react to, and spread information about, such events. Literature on viral information spread — as is the case with Internet memes — will also be covered, again with special emphasis on Twitter and related social media.

3.2.1 Twitter in Convergence Events: Activism and Democracy

First up in this section are qualitative and quantitative observations of Twitter microblogging in activism and democracy-building campaigns, or 'convergence' events as coined by Honeycutt and Herring [2009].

Two position papers presented at the *CHI 2010 Workshop on Microblogging* by Ems [2010] and Lin et al. [2010] provide a quick overview of the effectiveness of microblogging, particularly Twitter, in activism. Ems [2010] posited that Twitter is an effective tool to communicate and disseminate information by people in authoritarian regimes; and acts as a "sieve for news media outlets" by linking to other forms of media and "amplifying the distribution of other facts" for the knowledge of others [Ems, 2010].

The position paper by Lin et al. [2010] has a focus toward activism for disability awareness [Lin et al., 2010]: Twitter is found to efficient in promoting awareness by allowing the exchange of information, by allowing people to 'be heard', building an informal social network among supporters, and allowing for a viral spread of information (or 'marketing' a particular campaign).

Twitter was also observed to be a source of updates on the current status of whistleblowing site WikiLeaks (Twitter account: `@WikiLeaks`) after the recent 2010 exposé of confidential diplomatic cables[4].

By observing recent developments in the 2009 Iran Election (dissemination of pictures, video, and stories for public awareness), 2009 Moldovan 'Twitter Revolution' (to facilitate organization of protests), and the G20 Summit (in the role of 'informer' to help protesters avoid the police), Ems [2010] concurred that Twitter can be a threat to authoritarian regimes, providing power to the people by helping to shape public opinion. Such implications of microblogging in political and civil activism have been echoed in related literature, including:

[4] This announcement over WikiLeaks' official Twitter account is still accessible, at time of writing: <`https://twitter.com/wikileaks/status/6564225640042499`>

- **2009 Iranian election controversy**: Burns and Eltham [2009] performed a sociological evaluation of Twitter in the perspective of both the citizen protesters and the government. The early adopters of Twitter in this situation, by leveraging Twitter to spread awareness of the situation, was able to take the issue mainstream and reaching critical mass. Real-time broadcasts of updates and an online 'green campaign' (where supporters change the color of their Twitter profile picture to green in solidarity) was able to generate awareness of the situation in Iran to a wider audience despite threats of censorship. The spreading of the video of the shooting of a young protester, Neda, allowed the "protests [to gain] a broader, sympathetic audience" [Burns and Eltham, 2009]. However, Twitter inadvertently "became a vector for state repression", where it was used by the Iranian Revolutionary Guard and paramilitary to "hunt down and target Iranian pro-democracy activists" [Burns and Eltham, 2009].

- **2009 German elections**: Jungherr [2010] has observed the use of Twitter in the German *Superwahljahr 2009* and noted a 'rapid adoption' of microblogging use by politicians, parties, campaigns and supporters. It is used mainly as a tool for community building via Twitter account 'hubs' that form a focal point of discussion, e.g. the Twitter account `@teamdeutschland`, and a distribution channel for "social objects" [Jungherr, 2010]. Similar to prior findings on US political conventions [Hughes and Palen, 2009], Twitter is identified as a back-channel for communication in the German election context [Jungherr, 2010].

- **2009 Moldovan 'Twitter Revolution'**: Serbanuta et al.'s preliminary study [Serbanuta et al., 2010], although still a work in progress, has discovered several unique characteristics of Twitter activity pertaining to this event. An analysis of approximately ~28.5k tweets from ~1900 users (Twitter hashtag: `#pman` describing the event) provides quantitative statistics which allows further research into Twitter activity rate during such an political event: 14.8 messages per unique user contributing to the chatter, where ~10.7% of tweets contained links to other content, and tweets in Romanian and English were retweeted 2.6 times on average [Serbanuta et al., 2010].

3.2.2 Twitter's Role in Crisis Events

From the examination of the literature regarding *convergence* events, I will now explore work related to *crisis* events [Hughes and Palen, 2009], consisting of emergencies, disasters, and acts of God.

Reactions To Disaster

Sutton et al. [2008] were the first to apply studies of *back-channel communication* — defined as "public peer-to-peer communication" — in analyzing the use of social media during the 2007 Southern California Wildfires. Their findings pinpointed the growth and efficacy of social media channels in disaster and crisis response by the general public.

The idea of back-channel communication was expanded upon by Hughes and Palen [2009] in studying the adoption and use of Twitter in “mass convergence and emergency events”. By studying the Twitter activity in the Democratic and Republican National Conventions (which are political convergence events), and hurricanes *Gustav* and *Ike* in 2008, they observed that the usage patterns of Twitter during such events and the type of information shared in the form of tweets.

Hughes and Palen [2009] studied the message domain to isolate posts discussing the aforementioned mass convergence and emergency events; this allowed them to determine the type of information shared. They observed a prevalence of reply tweets and URL sharing in tweets, indicating the notion of information sharing and interpersonal communication (Section 3.5). Follow up research by Starbird et al. [2010] on the Canadian Red River Valley floods of 2009 studied the ‘social life’ of Twitter messages and the self-organizing behavior exhibited by users discussing the floods. “Commentary and the sharing of higher-level information” [Starbird et al., 2010], reply and URL sharing behavior [Hughes and Palen, 2009], sharing of experiences among flood survivors, and combination of tweets with authoritative news sources [Starbird et al., 2010] are exhibited in their research sample. In the user domain, Hughes and Palen [2009] determined that new users joining Twitter in the wake of mass convergence and emergency events tend to adopt Twitter use in the long term, in contrast with the general population on Twitter.

Vieweg and Starbird [2010], expanding on their earlier paper [Starbird et al., 2010], presented a position paper on analysis methods and challenges for microblogging in mass emergency events. They describe several types of future research that can be conducted on Twitter data in mass emergency: uncovering geo-location and geo-referencing information, studying retweeting of data as information ‘churn’ and at the same time an ‘informal recommender’ of timely crisis information, and also understand the influence of Twitter user network connections on their “message content, user stream behavior, and information spread during crises.” [Vieweg and Starbird, 2010].

On the other hand, Kireyev et al. [2009] experimented on using topic models on the message domain to classify microblog chatter during disasters, with unique challenges such as esoteric microblogging language patterns, short message length, and locale-specific text. As their work was still experimental, no conclusive results were reported.

Longueville et al. [2009] introduced the approach of mining spatio-temporal data from tweets to track forest fires in Marseilles, France. This is not dissimilar to the earlier study on river flooding by Starbird et al. [2010]. They obtain a data set of 313 tweets from 127 unique users during the 22 July 2009 Marseilles Fires. One of the challenges faced by the authors is that there is a low proportion of tweets from French users on Twitter, which made the findings in Longueville et al. [2009] relevant in a worldwide context. Temporal data were harvested from message timestamps, while spatial data were obtained from implicit place names or GPS coordinates found in a tweet (or failing which, the user profile of a tweet’s author). Four main research questions have been posed, in which qualitative data analyses are applied to answer each of them [Longueville et al., 2009]:

1. **Twitter is an extremely fast platform for information dissemination to report exceptional events**: the timeline of tweets rather accurately matches the real-world spread of the fire, with the exception of 'lag time' at the beginning of the fire.

2. **Twitter will provide accurate and useful spatio-temporal information** [as a location based social network]: location indicators such as place name mentions, hashtags of places, quantitative measurement of location/area, and user-positioning via GPS coordinates validates this hypothesis.

3. **Twitter users communicate with each other in widely open conversation; as a result, it is a primary source of information from citizens**: due to the combined presence of primary information (citizen journalism), and secondary information (aggregated data from RSS feeds with no added value) within tweets, this hypothesis cannot be confirmed as it stands. However, there is evidence that citizens exhibit self-organizing behavior when tweeting during emergency situations, as they quickly come to the acceptance of using unique hashtags to group their conversations in context, and center their conversations on these mutually-agreed-upon hashtags.

4. **Twitter is used as an information broadcasting and brokerage platform during crisis events**: up to 75% of tweets contain URLs with links to news media, and "dozens of pictures of the fires [were] taken and published" [Longueville et al., 2009] supports this hypothesis.

Another study exploring the usage of Twitter users in an emergency situation — this time on the Chilean earthquake of late February 2010 — was performed by Mendoza et al. [2010]. The study focused on tweets in the context of the Chilean earthquake, identified using the hashtag `#terremotochile` ('Chilean earthquake'), with the threefold objectives of: observing dynamics of news propagation and friend/follow patterns, influence of top users in the discussion, and to distinguish rumors versus actual news spread of a disaster on Twitter. Their dataset on Santiago, Chile-based users (based on timezone) contained 716k users with approximately 4.7 million tweets. Vital statistics from their observation on the tweet/user dataset are listed in Table 3.3.

With regards to their research foci, Mendoza et al. [2010] observed that Twitter discussion habits on the earthquake mirror the real world significance of the event in the real world: tweets of the disaster outnumber those discussing a popular Chilean music festival at that time. In fact, when keywords are grouped in term clouds, the temporal distribution of event words by day has a high correlation with the changing real-life events as a result of the earthquake; e.g. "tsunami" which reported on events on the first day, followed by "missing people" on the next day as a consequence. Retweeting behavior exhibited by users are tree-like, similar to [Kwak et al., 2010].

One of the discoveries in Mendoza et al. [2010] that can prove beneficial in future studies is the authors' quantitative measurement of the proportion of *truthful tweets* [*sic*] (i.e.

Feature	**Statistics**
Replies in tweets	98% of the tweets are reply-based
Follower-to-friend ratio	Almost half the users: more followers than friends (Authority users have >100k followers, e.g. CNN.)
User activity distribution	~64% of the surveyed users only wrote one tweet 11.47% of the users have more than 10 tweets (The rest are non-uniformly distributed in between)
Top user characteristics	Users with >2000 followers/friends: total tweets higher by one order of magnitude, compared with other users (Top 20 users during the event: news media, celebrities, non-profit organizations)
Effect of trending topics	Fraction of users writing about trending topics during the event: insignificant

Table 3.3: Mendoza et al. [2010]: Observations from their dataset of Twitter activity during the Chilean earthquake, February 2010.

substantiated facts) versus number of *rumor tweets* (i.e. unsubstantiated or false information). Out of a sampling of approximately ~4000 tweets, Mendoza et al. [2010] identified seven true stories and seven false stories. Each story is significant in that it has more than a thousand tweets in the original unsampled dataset. These stories are compared to external reliable sources (ground truth): over 95.5% of the tweets on confirmed events (the set of seven true stories) validate their truth. Conversely, 50% of tweets are observed to refute a story (from the seven false stories) when it is evidently false. These findings suggest that "the Twitter community works like a collaborative filter of information" [Mendoza et al., 2010].

The studies by Longueville et al. [2009] and Mendoza et al. [2010] complement existing research on crisis events as they help researchers understand the role of microblogging in disaster events from a multicultural perspective, due to the difference in usage habits on Twitter and tweet languages across countries.

Early Detection and Warning of Disaster

Another emerging aspect of research on Twitter use in emergencies is the early detection and warning of potential emergency situations, and to complement existing sensory and surveillance systems.

Sakaki et al. [2010] suggested the use of real-time 'social sensors' in earthquake detection, as Twitter users frequently post details of earthquakes and tremors on Twitter as soon as they feel the event happening. They propose an event detection methodology to validate findings of earthquakes, using Support Vector Machines (SVM) to classify earthquake-related messages with three features: "number of words in a tweet message and the position of the query word within a tweet; the [keywords] in a tweet; [and] the words before and after the query word" [Sakaki et al., 2010]. Such processing is done on Japanese language tweets. The authors propose another probabilistic model to handle spatio-temporal information about tweets using Kalman and particle filters to predict the earthquakes' trajectory, which is then used to create a prototype earthquake reporting system [Sakaki et al., 2010].

Finally, researchers for the US Geological Society [Guy et al., 2010], developed a *Twitter Earthquake Detector (TED)* for earthquakes globally, similar to the Sakaki et al. [2010] as they used both the Twitter user and message domains to do so. By 'listening' to the Twitter Streaming API for incoming tweets mentioning earthquake-related terms in several languages, matching tweets are then dumped into a database and sanitized to remove instances of retweets and aggregator users [Guy et al., 2010]. Similar to other work — such as my prototype Cheong and Lee [2010c], detailed in Section 5.1 — the authors used the Google Maps API Geocoder Service to find location data in tweets, such as profile location or GPS coordinates. By using a mathematical model to find the earthquakes' probable epicenter from the spatial and temporal information found in the collected tweets, and another model dealing with the significance of an earthquake based on user activity, the researchers are able to come up with a report and graphical map overlay detailing findings of a quake, which complements the traditional geo-monitoring data [Guy et al., 2010].

However, there are several known issues identified within the system by Guy et al. [2010], i.e.:

- **Completeness**: The lack of geo-location information in the content/metadata of many tweets, which require a high number of tweet samples to fix.

- **Ambiguity**: The nouns used to detect quakes are ambiguous; e.g. the word *quake* itself can refer to both an earthquake and the computer game *Quake*.

- **Delay**: Twitter activity spikes that occur only after the real-world quake events have elapsed.

Despite that, the authors conclude the paper with several interesting findings. The Twitter chatter on a quake outperforms traditional sensors in two instances. One, a Melbourne quake with a low-enough Richter magnitude has been detected in a timely manner; two, TED performed four times faster than traditional detectors for another earthquake in Indonesia [Guy et al., 2010]. Despite the minor shortcomings, time series comparison revealed that peaks in Twitter activity correlates with actual quakes, and the signal-to-noise ratio of earthquake tweets is high enough to warrant effective detection [Guy et al., 2010].

3.2.3 Viral Information Spread and Memetics

The spread of viral information on the Internet, particularly in the form of *memes*, have been documented as far back as 2000.

An *Internet meme*, nowadays shortened to just *meme*, has its etymological origins in Dawkins [1989], where it is defined as "...a noun that conveys the idea of a unit of cultural transmission" [Dawkins, 1989]. In its current connotation (on the Internet), memes generally refer to pieces of information — joke, captioned image ('image macro'), and funny videos on YouTube being good examples — that spread from user to user. This is not dissimilar to how emails have been forwarded from person to person in the early days of the Internet [Flor, 2000].

Memes that spread across the Internet (originally via email, as mentioned) have been known to exist before the days of Web 2.0 social media. Examples of Internet memes being documented in literature include [Hodge, 2000; Flor, 2000].

Back to current research, a study on memetic spread performed by Arbesman [2004] on blogs show that the spread of viral information can be observed in terms of relative 'spikes' or increases in traffic to the memetic blog in question. This experiment illustrates the power of the Internet in its ability to spread information similar to an 'epidemic' [Arbesman, 2004].

Wasik [2009] has documented experiments on memetic spread of information (via email, blogs and the Web) conducted from 2003-2004. The memetic spread of ideas in collective action (flash mobs), entertainment, politics, and corporate marketing are revealed in a series of experiments which are well-documented. From popular science, a parallel can be drawn from the *tipping points* idea by Gladwell [2002], and the *wisdom of crowds* theory by Surowiecki [2005] which illustrate how collective coordination helps spread an idea until it reaches *critical mass.*

Research on information brokering, diffusion, and viral spread with emphasis on microblogging (specifically Twitter) has since become available in 2010. Van Liere [2010] studied patterns of information brokering and geographic diffusion of retweets on Twitter by first proposing three patterns: *uniform distribution*; a *local pattern* skewed towards nearby users; and the *information brokerage* pattern due to like-interests skewed towards users furthest away.

Van Liere [2010] first scanned all retweeted messages before sanitizing of inaccessible URLs and skewed data, then used the conversation topic as a search key, in order to retrieve all mentioned tweets. By obtaining user profile information on geographic coordinates for each of the authors, Van Liere [2010] managed to determine the exact location and is able to perform a Haversine distance calculation on the exact geographical distance between users. From a dataset of 6,424 geocoded users, the author obtained ~13.4k retweets stemming from 285 original posts [van Liere, 2010]. Temporally, about 60% of retweets commence in the first hour after the original tweet was posted, and this quickly fades to zero in the time period of about a day. Geographically, he determined that the average retweet distance is approximately ~955km, with the median being ~1698km, suggesting an *information brokerage* pattern [van Liere, 2010].

The paper by Van Liere [2010] also touched on the three-fold motivations behind a retweet: to vie for attention and increased follower count; to gain influence as a social media filter "who specializes in a particular topic"; and to transfer information from one social network group to another - acting as a bridge between distinct groups. In fact, the first two motivations mentioned are complemented by Metaxas and Mustafaraj [2010]:

"...one is much more likely to retweet a message coming from an original sender with whom one agrees... [or] shares political orientation [with]", as the majority of users "were very unlikely to retweet a message that they did not agree with" [Metaxas and Mustafaraj, 2010].

Abrol and Khan [2010], by observing frequency of tweets mentioning a specific location versus the actual real-world population of the location, was able to come up with a metric called the *Frequency-Population Ratio (FPR)*. This can then be used as a yardstick to measure any anomalies with respect to patterns of information spread involving issues regarding a particular location. For instance, the Fort Hood shooting in November 2009 had an abnormal FPR (over 1800 compared to the baseline average of 1), indicating that information on some event happening in that particular place has exhibited signs on spreading rather rapidly.

Finally, Stonedahl et al. [2010] measured viral marketing strategies on four theoretical models alongside the real-world case of Twitter, which would detail the viral nature of information spread. The user network on Twitter was crawled, starting with a random seed Twitter user, using breadth-first search; 999 closest nodes from the seed and all 13,343 reciprocal friendship links between them, as was investigated by Kwak et al. [2010]. The findings that set Twitter apart from other theoretical models illustrated that Twitter does not take on the form of a random social network; in fact, it has hubs of users with high degree of friends situated close together in the social graph [Stonedahl et al., 2010]. Bridges between unconnected parts of the network (information brokers) are situated rather far away; leading them to conclude that Twitter is substantially different to commonly-found networks in social network literature, including the regular *lattice* network and the *small-world* network [Stonedahl et al., 2010].

Revisiting [Schafer, 2010; Java et al., 2009] which have discussed earlier (Section 3.1.2), Twitter's distribution of friends and followers in the user network seems to obey a power law. This is indicative of a *scale-free network* [Barabási and Albert, 1999], where a power-law distribution governs the number of edges connected to a node [Barabási and Albert, 1999]: in this case, within the Twitter social graph. Scale-free networks are also inherent in e.g. the topology of the World Wide Web, where nodes in the network consist of web pages, and the edges are the hyperlinks between such web pages [Barabási and Albert, 1999].

3.3 Emergent Behavior and Pattern Recognition

Having reviewed existing research on the spread and 'social life' of information through Twitter and related social media technologies in the previous section, attention is now given to the third theme of research that is the focus of this literature review.

One can observe emergent properties on Twitter, owing to the myriad of complex interactions taking place between users through the *following* mechanism, and the generation of volumes of tweets in their daily usage of Twitter. Thus, this section surveys the literature on the kinds of emergent patterns that are visible on Twitter, which commonly take the form of usage spikes and idiosyncratic patterns amongst groups of users (and their tweets).

3.3.1 Trend Analysis: Blogs and Social Media

Event detection and trend analysis are common research areas in blogs and other social media that are of interest in the study of Twitter and microblogging.

I will now review the literature on the usage of online social media (including blogs) to mirror real world happenings. Gruhl et al. [2004] studied blogs in an attempt to predict real-world rankings of books on Amazon.com based on the volume of chatter generated in the *blogosphere*. They found a correlation between the 'buzz' generated for a particular book with the real-world sales performance on Amazon.com in terms of sales rankings. Their work also introduced a novel concept of *spike prediction* based on the knowledge of existing blog chatter of a particular book [Gruhl et al., 2004].

Choudhury et al. [2008] performed a similar study in 2008 to correlate blog communication dynamics on particular stock counters with their real-world performance in the stock market. Their framework managed to map the behavior of blog commentary with the real-world performance of a particular stock, with a low error rate. The papers by Gruhl et al. [2004] and Choudhury et al. [2008] suggest that online chatter on blogs can 'mirror' real-world happenings rather efficiently.

Fukuhara et al. [2005] also found a link between blog articles with real-world temporal data, where mentions of topics in the Japanese blogosphere are found to have a connection to real-world social events, weather, and topics reported in the Japanese mass media. Gruhl et al. [2004] also concluded that *spike topics* from real-world events can affect spiking behavior in blog postings.

To detect anomalous users in Twitter and microblogging, related research with dynamics of phone networks by Gupta and Dey [2010] can potentially be applied to the user and message domains. They use a set of feature vectors that, for both incoming and outgoing communications, identify both the degree of communication and the length of communication; e.g. $f_{OUT}(low, short)$ stands for count of outgoing messages to low-frequency contacts where messages are short. This degree of communication metric is coupled with a set of global features such as the sum of contacts with one-way activity, and total time period. Using a k-nearest neighbor classifier, they successfully applied their methodology on the Enron email dataset and the IEEE VAST 2008 challenge dataset [Gupta and Dey, 2010]. Gupta and Dey [pers. comm., 23 August 2010] opine that such research has potential to be adapted for Twitter, in terms of its message domain (for tweet length) and user domain (friend/follow activity).

3.3.2 Trend and Anomaly Analysis from Twitter

Similar to the previous section, the literature on trend analysis (and anomaly analysis) with a specific focus on Twitter — hitherto only found in blogs and other social media — have increased in number since 2010.

Kumar et al. [2010] performed an experiment to mathematically model the dynamics of conversations (messages and users) in online social networks. One of their findings is that Twitter hashtags exhibit distinctive behavior when they are trending:

- **Memes**: Topics with a high 'preferential attachment' ("messages that have already received many replies [which] are more likely to receive a new reply" [Kumar et al., 2010]) tend to be memes.

- **Current events**: Topics with a high 'copying rate' ("new authors tended to join in often" [Kumar et al., 2010]) have often a stronger 'sense of time', and mainly consist of current events.

Revisiting the work of Kwak et al. [2010] (earlier discussed in Section 3.1.2), which also contained research on trends and Twitter trending topics. Trending behavior exhibited by different trending topics are different, despite their similarities in the total number of tweets recorded [Kwak et al., 2010]. In fact most trending topics have mostly one 'spike' (73% of cases) compared to multiple spikes; about 31% lasted one day or less, but only about 7% persisted for more than ten days [Kwak et al., 2010].

A readily-available example is my work on such topics, located in Section 6.3 and published in [Cheong and Lee, 2010b], illustrates the point by Kwak et al. [2010]. I have studied two trending topics: the Apple iPhone 3 launch, and the 2009 Iran Election controversy. The *iPhone* topic had more users taking part in discussion compared to the *Iran Election* topic; however, the pace of discussion slowed down quicker than the latter [Cheong and Lee, 2010b].

Therefore, Kwak et al. [2010] came up with a new, two-attribute, classification scheme to categorize trending topics. *Criticality* is defined in terms of its potential to spread, and *exogeny/endogeny* refers to external versus internal "factors that push a topic to the top trending topic list and [cause] the spread of the topic" [Kwak et al., 2010]. Four categories were derived from the two sets of two attributes:

- **Exogenous critical**: consists of timely breaking news topics such as about celebrities, causing a single spike but with gradual decay.

- **Exogenous sub-critical**: consists of 'ephemeral' micro-memes and hashtags, usually causing one spike and rapid decaying over time.

- **Endogenous critical**: consists of topics of a "more lasting nature [such as] professional sports teams, cities, and brands... [labeled as] persistent news"[Kwak et al., 2010], that have multiple spikes and showing signs of slow decay in activity.

- **Endogenous sub-critical**: same as endogenous critical, but the trending period is much shorter.

It is pertinent to note here that in the early stages of research, I have proposed a similar trend categorization theme predating Kwak et al. [2010]; where I categorize a trend based on its temporal distribution of tweets. This was published as [Cheong and Lee, 2009], and documented in Section 8.2 later in this thesis. Work by Kwak et al. [2010] above has independently corroborated the validity of my findings.

Huang et al. [2010] also conducted studies on trends from a hashtag perspective, which reveals more patterns on user activity contributing to tweets over time. By obtaining the standard deviation between time-stamps of messages of a single hashtag, they found that a small standard deviation of time-stamps indicated '*conversation*', or micro-memes "...both adopted and abandoned in a short period of time" [Huang et al., 2010]. This is in contrast with message sets with large standard deviations of time-stamps, indicating '*organization*', which involved serious topics such as news on current affairs.

The authors also observed the "skew of the tag time-stamps" [Huang et al., 2010] in the time-series graph of a tag to measure the viral nature of said tag. A negative (left-sided) skew indicates gradual adoption of a tag before reaching its peak of activity, compared to a positive (right-sided) skew which indicates a rapid adoption of a tag before its gradual decline, as evident in micro-memes [Huang et al., 2010]. Kurtosis, or the fourth moment, represents the "staying power" of a hashtag [Huang et al., 2010]; topics with high kurtosis represents bursts of temporal activity (such as trending micro-memes which later fade to zero), and a low kurtosis indicates a persistent topic (e.g. the long-discussed H1N1 flu outbreak). As such, the kurtosis metric "can be used to differentiate between micro-memes, recurring tags, or spam" [Huang et al., 2010].

3.3.3 Pattern Detection and User Clustering on Twitter

More studies have been performed on pattern detection and user clustering based on exhibited demographic and messaging patterns by Twitter users. This subsection discusses two such studies: one newly developed after the introduction of the *Lists* feature on Twitter [Kim, Jo, Moon and Oh, 2010]. *Lists* allow users to group their friends according to custom categories. The second study, on the other hand, is a thesis on the automatic classification of tweets [Horn, 2010].

Kim, Jo, Moon and Oh [2010] have turned Twitter *Lists* into an invaluable source for detecting commonalities among users in the Twitter community, using information from the user and message domains combined. They posit that *lists* — a publicly available data source on Twitter (Section 4.1.4) — have implicit characteristics for the modeling of commonalities between users, as judged by their peers. The usage of lists mean that keywords do not have to be explicitly mentioned in tweets; by merely looking at a user's association, one can deduce the degree of commonality between herself and her peers.

For their experiment, Kim, Jo, Moon and Oh [2010] used a dataset of ~3.3 million users (approximately ~10% of the entire user base of Twitter) belonging to ~900k lists. Similar list names were combined together in groups (about 2-3 per group). Chi-square feature selection is applied to the corpus of all tweets belonging to each single user, and repeated for all users belonging to a single list group. To obtain the ground truth, human experimenters associate Twitter users with a particular keyword that best describes the individual user [Kim, Jo, Moon and Oh, 2010]. The list of high chi-square words — terms with the highest relevance to a list that distinguishes it from the rest — are then compared against the words picked by the human experimenters; resulting in an accuracy rate of 0.925. The authors conclude their experiment by stating that “the combination of the Twitter list functionality and the chi-squared feature selection is an efficient tool for inferring user characteristics” [Kim, Jo, Moon and Oh, 2010]. The aforementioned combination could potentially be augmented with profile information and friend/follow characteristics of the users in future research.

Horn [2010] contributed to the body of knowledge with his Masters’ thesis on automatically classifying tweets — and by extension users — with potential applications of separating user-generated content from professional content and spam filtering[5]. He proposed two mutually-independent, separate, classification schemes:

1. **C1** distinguishes between user types, and has the following classification categories: ‘news’; ‘users’ describing everyday users of Twitter; and ‘company’ containing company promotions, inclusive of spam and sponsored tweets.

2. **C2** distinguishes between pure ‘facts’ and ‘opinions’ consisting of subjective user tweets, quotations, or questions.

For his thesis, Horn [2010] created a data set of ~4,800 tweets from 120 users; and for supervised clustering, chose ‘news’ users from Twitter accounts of well-known news corporations/newspapers, crawled ‘regular users’ from several celebrity users as seeds, and highlighted ‘company’ users from scam websites and spam keywords.

Several findings from analysis in the *C1* category showed that the statistical figures — typical number of distinct words used (and also the rankings of keywords), average length of a tweet, and average interval between tweets — are clearly different between samples from all three types of users. The same finding applies to analysis of the two different tweet types (factual versus opinionated) in *C2*: factual tweets have almost double the number of distinct keywords compared to opinion tweets, and the average time between tweets are almost ten times higher for factual tweets. Different levels of sentiment are also found for each separate category (such research is further described in Section 3.4).

Sriram et al. [2010] has identified eight features (‘*8F*’) of automatically classifying tweet texts to one of the five categories of news, events, opinions, deals, and private messages. Using the ‘bag-of-words’ concept as a comparison, they applied the Naïve Bayes classification method [Hall et al., 2009] on the eight features on 5407 tweets from 684 authors. Human experimenters provided the ground truth by manually assigning a category to each tweet. The *8F*s consisted of [Sriram et al., 2010]:

1. author name;

2. presence of Internet shortenings, emotions, and slang words;

3. presence of time-event keywords (e.g. participant, place, and time information);

[5] *Spam users* on Twitter are defined as users with aggressive *following/un-following* behavior not commonly found among normal Twitter users; and *spam* on Twitter are links to phishing and malware sites and unsolicited advertisements.

4. presence of opinioned words;
5. presence of emphasis by capital letters or repeated syllables (e.g. `veeeery`);
6. presence of currency (`$`) or percentage signs (`%`);
7. presence of `@user` notations at the beginning of the message;
8. presence of `@user` notations within the message.

Sriram et al. [2010] concluded their study by stating that their *8F* classification performs better than traditional bag-of-words classification to "...classify tweets into general but important categories by using the author information and features within the tweets" [Sriram et al., 2010].

The studies by Sriram et al. [2010] and Horn [2010] dealt primarily with the message domain on Twitter, with the author information as the only exception from the user domain. However, the findings demonstrate the ability of using tweets to indirectly classify users; therefore I posit that the incorporation of features from both users and messages will increase the accuracy of future research on pattern detection on Twitter.

3.3.4 Pattern Detection for Twitter Spam Detection

A novel area of investigation based on pattern detection in user behavior and their message characteristics is the application on spam detection and filtering. Moh and Murmann [2010] and Lee et al. [2010] have both performed studies on this, and their findings based on analysis of the friend/follow patterns and live 'honeypot' analysis will be detailed here. Moh and Murmann [2010] adapted existing features found from the Twitter users API and synthesized few new attributes/metrics that have been useful in spam detection:

- **Friend/follower count**: derived average friends per day, average followers per day, percent of reciprocal friends
- **Total favorites count**
- **Protected updates flag**
- **Status update count**: derived average updates per day
- **Presence of profile URL**
- **Username**: derived presence of numbers in username string
- **Trust metric**: the sum of (1/users followed) for all followers of a given user

They use the metrics above not only for individual users, but also for all peers of a given user (a peer is defined as a reciprocal friend, cf. Kwak et al [Kwak et al., 2010], who has a mutual friend/follow connection). From their findings after applying supervised learning algorithms — *JRIP*, *J48*, *SOM*, and *Naïve Bayes* — as provided by the Weka data-mining package — it is evident that real spammers (as validated by human experimenters) have the following obvious features that discriminate them from the rest, listed by order of decreasing importance [Moh and Murmann, 2010]:

- ratio of spammers to legitimate followers;
- average friends per day;
- trust metric (and several weighted variants);
- friend to follower ratio;
- friends' average friend to follower ratio; and
- followers' average of protected users.

In a similar vein, Lee et al. [2010] have identified 14 classes of attributes from both the user and message domain, of which the ones not covered above by Moh and Murmann [2010] are:

- **Account age**: the duration of time a user's Twitter account has existed for;
- **URLs in the message content**: the ratio of URL counts over total tweet count, and ratio of unique (non-duplicate) URLs over total tweet count were derived;
- **`@user` mentions in message content**: the ratio of `@user` mentions in the last 20 tweets, and ratio of unique `@user` mentions in the last 20 tweets were derived; and
- **Average content similarity metric**: the standard cosine similarity over the bag-of-words vectors (tweet content) was used.

Abrol and Khan [2010], in the context of their Twitter geo-content study (discussed in Section 3.2.3), came up with a simple formula to deduce spammers based on the fact that their behavior in the user and message domains are atypical of a normal user. As above, the ratio of followers-to-friends is small; another finding is that a spammer "rarely addresses his messages to some specific people" [Abrol and Khan, 2010] using the `@user` or `RT` notations. Based on these two findings, for a given message on Twitter, they proposed a formula which "...tags to a certain level of confidence whether the message is spam or not", based on properties of the message's author:

$$\text{Spam confidence} = \frac{1}{\frac{\Sigma\text{followers}}{\Sigma\text{followees}} + \mu\left(\frac{\Sigma\text{reply tweets}}{\Sigma\text{tweets}}\right)} \tag{3.1}$$

Qualitative observations on spammers and spam users in the Twitter environment have also been performed, both in a honeypot study by Lee et al. [2010], and also a study on political opinion-spam by Metaxas and Mustafaraj [2010]. The honeypots by Lee et al. [2010] observed that 'social spammers' frequently distribute malware, spam, phishing messages and affiliate programs through contextual 'social'-based spam messages; the following *modus operandi* is observed:

- **Duplicated tweets**: spam tweets were duplicated verbatim to multiple unsuspecting users;
- **Pornographic references**: found in spam profiles and tweets;

- **Mixed content**: spam users mix legitimate content with advertising, promotional, phishing, or spam content;

- **Trust and user infiltration**: “infiltrators” behave like a normal user initially but then start disseminating spam after reaching a sizable number of followers.

Another comprehensive study by Thomas et al. [2011], combining qualitiative observations of real-world Twitter spammers with quantitative metrics recorded by such users, was conducted to evaluate the “...the tools, techniques, and support infrastructure [Twitter spammers] rely upon” [Thomas et al., 2011]. The authors studied “over 1.1 million accounts suspended by Twitter” [Thomas et al., 2011] as said accounts were identified to be spam users. Several key findings from [Thomas et al., 2011] are:

- **Common methods of spamming detected by Twitter**: the modus operandi of the majority of spammers include “...frequent requests to befriend users in a short period, reposting duplicate content across multiple accounts, sending unsolicited mentions, posting only URLs, and posting irrelevant or misleading content to trending topics” [Thomas et al., 2011]. Accounts exhibiting such activities are frequently suspended by Twitter Inc. for violation of terms of service.

- **Dichotomy in tweeting behavior by spammers**: [Thomas et al., 2011] detected two kinds of spammers on Twitter. The first kind consists of short-lived Twitter spam accounts which go all-out on spamming as much as possible before being suspended. The second kind generates spam tweets infrequently, but over a longer period of time. (Another group of spammers work within the allowed limits, but are banned by Twitter for different reasons).

- **Follower-friend disparity**: Spam Twitter users find it difficult to form ‘social relationships’ with other Twitter users for obvious reasons. However, some spammers who do so have a disproportionate number of people followed (friends) compared to followers, “...indicating a lack of reciprocated relationships” [Thomas et al., 2011].

- **Spam URLs**: The URLs being pushed to other users by Twitter spammers typically consist of link shorteners such as *bit.ly* or free custom subdomains such as *dot.tk* to mask the true Web address of their product. Such dodgy ‘products’ consist of advertising revenue-generating schemes that only serve to enrich the spammers [Thomas et al., 2011].

Meanwhile, Metaxas and Mustafaraj [2010] observed real-world political opinion-spam disseminated through aggressive campaigns, with a similar modus operandi to Lee et al. [2010]:

- **Low connectivity**: spammers tend to be unconnected in the friend/follow social graph, ompared to other ‘political tweeters’ who have similarities of user profiles/tweet sentiments with others of the same ideology;

- **Fake users**: spammers create bogus user accounts relating to the topic being discussed; e.g. spam links with text
  `coakleysaidit`[6] from 9 spam users sent 929 tweets addressed to 573 unique users in only approximately two hours;

- **Targeted content**: spam accounts target users with certain message content.

3.4 Modeling, Sentiment Detection, and User Search

The current section of the literature survey is dedicated to the fourth theme of research pertaining to Twitter, which pertains to both the user and message domains. This theme encompasses applications in the fields of: information retrieval, personalization, opinion and sentiment analysis. These research areas are by no means new; the novelty comes from the fact that these research topics are applied directly on Twitter users and messages.

3.4.1 User Ranking and Topic Modeling

Ritter et al. [2010] performed research on unsupervised modeling of Twitter conversations to detect dialog structure between users. Using ∼1.3 million conversations from the Twitter public timeline, their first discovery is that each conversation has a range of between 2–243 posts. Then, they sanitize their data by removing non-English tweets, before clustering together misspelled words using the *JCluster* algorithm. Conversation and topic modeling, employing the Latent Dirichlet Allocation (LDA) and Bayesian methods, is performed on the resulting message set. The authors found that coherent patterns do emerge from messages between users on Twitter, and that unsupervised modeling of dialogs on Twitter is indeed possible even with a large dataset. As a result, the authors also come up with a ten-act conversation and topic model corresponding to different features e.g. reactions versus questions (discussed in further detail in Section 3.5).

Puniyani et al. [2010] conducted research in a similar vein. They construct an LDA [Blei et al., 2003] topic model over ∼837k Twitter messages to identify latent themes within. Due to the small size of a Twitter message, the approach in Puniyani et al. [2010] aggregates all messages from a user into one 'document', learns the latent topics that characterize authors, and infers the underlying user network structure, user similarity, and connections between users with similar latent topics. As the experiments were still in progress as of time of writing, their findings are still inconclusive [Puniyani et al., 2010].

3.4.2 Recommender Systems and Personalization

Bernstein et al. [2010], researching on users' need to customize their user streams, found out that at least a thousand tweets can be received and consumed daily by an active user. By performing a user study among 78 users, they found out that the average number of tweets daily among this group is around 786 (standard deviation = 658), which is too

[6] Referring to the Coakley–Brown Senate race, January 2010: <`http://en.wikipedia.org/wiki/United_States_Senate_special_election_in_Massachusetts,_2010`>

much for a user to consume per day [Bernstein et al., 2010]. From the survey, 70 users categorized their Twitter use as personal, while 16 classify them as professional, indicating an overlap of Twitter usage for both personal and professional purposes. Bernstein et al. [2010] have also proposed three ideas: users will value tweets based on the relevance to their interests; *tie strength* i.e. users would want to see more from users they are connected to via Twitter; and *serendipity* i.e. random tweets which are interesting. Although still a work in progress, Bernstein et al. [2010] have prototyped a visual personal Twitter feed, which displays 'trends' custom-tailored to an individual user's preferences as a quick 'first-pass' filter to relevant topics of interest.

The *Buzzer* recommender system by Phelan et al. [2009] is a web-based interface that utilizes the *Lucene* information retrieval engine to mine the content of Twitter messages for topics of interest; allowing the user to view personalized streams either in the public timeline, friends' combined messages, or simple term-frequency from RSS feeds. Wu et al. [2010] has a similar research topic in using the Term Frequency-Inverse Document Frequency (*TF-IDF*) and *TextRank* mechanisms to perform automated tagging and annotation using keywords extracted and sanitized from a users' collection of Twitter messages.

An observation from this section is that current research surveyed primarily operate on the message domain, incorporating only the basic ideas from the user domain such as the basic follower network. Interestingly, findings from the message domain such as the tagging and annotation algorithm by Wu et al. [2010] does provide for the inference of user habits/behaviors, which would be useful in augmenting future user domain studies.

3.4.3 Opinion Mining and Sentiment Analysis

Research on the usage of Twitter for sentiment and opinion analysis have started becoming available toward the last quarter of 2009. Here, the state of the art in this particular field is highlighted.

Traditionally studied in the contexts of blogs, fora, and product reviews (cf. Pang and Lee [2008]), applications of opinion mining and sentimental analysis on microblog content, especially the message domain on Twitter, has been given increasing attention lately

Wasow and Baron [2010] position paper on using tweets as a measure of measuring user sentiment — themed the *Bruno Effect* — tries to study the link between microblog sentiment and the real-world 38% decline in revenue for the *Bruno* movie release in 2010. They hypothesize that comments in Twitter (which reflect user sentiment) is associated with real-world decline in sales performance for the said movie. Using the *TweetCritics* sentimental analysis tool for Twitter, they are able to come up with a model of the difference in revenue (between the first and second days) as a function of the sum of negative tweets recorded in the same time interval. Their secondary analysis reveals that automated sentimental analysis tools such as *TweetCritics* are comparable to manual sentimental analysis of tweets by human experimenters; Wasow and Baron [2010] used the online Amazon Mechanical Turk crowd-sourcing system.

An application of Twitter to analyze its users' opinions and to help annotate events was discovered by Shamma et al. [2009] who investigated Twitter chatter (the message domain) during the 2008 US Presidential Debates. Their first finding corroborates with literature on spike analysis (Section 3.3.1): they found that the “level of Twitter activity serves as a predictor of changes in topics in the media event” [Shamma et al., 2009]. Shamma et al. [2009] also delved in the user domain to map conversational structures between users, in terms of `@user` messages. The matching of context was possible between the actual texts of the debate obtained through close captioning with the Twitter chatter; viz. a form of opinion mining with regards to the collective reactions exhibited by Twitter users toward a particular topic debated [Shamma et al., 2009].

Shamma et al. [2010] have followed up on their earlier work [Shamma et al., 2009] with a position paper summarizing and supplementing their original findings on the US Presidential Debates; short tweets were also found to be effective in annotating debates based on popular opinion. The said annotation is made possible by the availability of temporal data (message time-stamps), and that statistics on Twitter messaging activity over a specified time interval can produce cues on the venue, structure, and the activity level in a real-world program [Shamma et al., 2010, 2009].

A preliminary short study by Pennacchiotti and Popescu [2010] on the usage of sentimental words and “controversial terms (e.g. `trial, apology`) from Wikipedia pages” [Pennacchiotti and Popescu, 2010] allowed them to develop an early prototype of a controversy detector for Twitter messages.

On the other hand, Bollen et al. [2009], using *Profile of Mood States* (POMS) analysis to study changes of sentiment on a corpus of 9.6 million tweets (from August–December 2008), have successfully used tweet-based sentiments to model real-life socioeconomic phenomena. Their list of such phenomena is adapted from a list of twenty real-life events, incorporating the corresponding stock market behavior of the *Dow Jones Industrial Average* and *West Texas Intermediate* oil price indices. They have found in the study of several events that changes in terms of ‘mood states’ were evident based on analysis of the sentiments exhibited in public tweets, such as:

- **US Presidential Election**: doubt before the election (increased mood states of confusion/depression), celebration after the election (increased mood state of vigor, decreased mood state of fatigue).
- **Thanksgiving holiday**: increased mood state of vigor, decreased mood state of fatigue.
- **Dow Jones Industrial Average price drop due to bailouts**: increased mood state of depression.
- **John McCain announcing Sarah Palin as his running mate for the elections**: increased mood state of tension.

Jansen et al. [2009a] have proposed an automated framework for brand sentiment analysis in terms of user sentiment in relation to certain products via Twitter messages. They used the *Summize* service, later acquired by Twitter as Twitter Search (Section 4.1.1) to analyze “tweet sentiment” [Jansen et al., 2009a] with regard to a set of nouns describing brand names. In the message domain, the intention is categorized into four classes: sentiment, information-seeking, information-providing, and comments [Jansen et al., 2009a]. The message length of a tweet is also taken into account to study the linguistics of the user base when commenting on a brand; co-occurrence of key terms and phrases such as personal prepositions are also identified in the messages. In the user domain, the authors tracked the volume of tweets between customers and the brand/company's Twitter account, and their frequency to determine communication patterns [Jansen et al., 2009a].

Certain co-occurrences of words found in tweets was observed to correlate to the sentiments or intentions of their authors [Jansen et al., 2009a]. About 19% of the total tweets "mention an organization or product brand in some way" [Jansen et al., 2009a] and about 20% of these tweets also "expressed a sentiment or opinion concerning that company, product, or service" [Jansen et al., 2009a]. The findings from Jansen et al. [2009a] indicated the suitability of Twitter as an avenue for mining sentiments of users.

Wrapping up this subsection, note that almost all the literature discussed in this subsection — on opinion mining and sentiment analysis — deal exclusively with the message domain on Twitter. This matches the traditional practice of using pure textual content for sentimental and opinion analysis. Again, suggestions for future research include combining such analyses with observations and measurements from the user domain.

3.4.4 User Search

User search, the focus of this subsection of the literature survey, analyzes research that deals with a user's experience while searching for tweets on Twitter. This relates to other topics in this section in the fact that search has traditionally been a problem area in information retrieval.

A position paper by Suh et al. [2010] on Twitter with regards to sense-making has, among other things, defined the need for search as an information-seeking task on Twitter. The limited length of messages is identified as a problem "for [conventional] search algorithms to work efficiently" [Suh et al., 2010]. Term expansion is suggested as an approach to handle situations involving Twitter search.

From a user study, Golovchinsky and Efron [2010] have also found that Twitter search is becoming more frequently used by users trying to locate information such as "[for] events... for people, and for trending topics" [Golovchinsky and Efron, 2010]. They also identified weaknesses from the Twitter API, previously discovered in e.g. Cheong and Lee [2009] and Kwak et al. [2010]. One such important weakness is the very limited window for search results, restricted to merely two weeks (Section 4.1.3). The usage of hashtags is said to improve search experience from the results of the user study [Golovchinsky and Efron, 2010].

Several caveats of Twitter search include the "small size of each tweet [making] it hard to estimate the relevance of that tweet" [Golovchinsky and Efron, 2010]; and the traditional Twitter search interface "...[makes] it difficult to extract key memes or documents that characterize an event" [Golovchinsky and Efron, 2010]. These points are also summarized by researchers such as Böhringer and Gluchowski [2009] who identified potential in the area of search and information retrieval; in the authors' words, searching "for a subjectively important information in the data stream" in the "proverbial haystack" [Böhringer and Gluchowski, 2009]

An experimental study on information retrieval by Sharifi et al. [2010] focused on the development of an automatic summarization system that takes trending topic keywords, and applies phrase reinforcement algorithms to automatically generate summaries of tweets discussing a particular topic.

3.5 Human Factors on Twitter

I shall now move on to the fifth theme of current literature studied in this chapter. In order to understand the rationale behind human interaction on Twitter (in the user domain), one needs to understand human factors, i.e. what makes humans communicate and interact with one another the way they do. Hence, this section covers four such major issues: parallels between Twitter and other communication channels; human 'presence' and sharing; privacy concerns and implications; and the need for social communication inherent in humans.

3.5.1 Similarities in Other Communication Channels and Social Networks

Research which will be covered in this subsection — relating to parallels between Twitter communication and other forms of traditional and modern communication — will provide an insight into the similar human factors involved, such as motivations and usage practices. Three of the five papers below in this section deal with cellular phone communication networks, which have existed for more than a few decades.

Communication Patterns

Although not strictly within the realm of Twitter, Dearman et al. [2008] have performed a field study on the information needs and sharing opportunities in normal everyday life. Research in Dearman et al. [2008] involved volunteers manually jotting down their "daily information needs and sharing desires" in a diary which is then analyzed by the authors. They have identified 9 distinct 'information categories' and 21 subcategories to classify a question or piece of shared information. For instance, the question "*Is it too cold outside to go running?*" [Dearman et al., 2008] fits into the usage intent of "asking a question" and also the information category of "environmental conditions", specifically the weather.

Rangaswamy et al. [2010] studied an Indian SMS service, *SMS Chatter* that parallels the structure of Twitter messages in terms of short message lengths, but is not an online social network *per se*. The basic premise of this service is that users subscribe to groups or *chat rooms* and send their messages to subscribers within the same group — not unlike Internet Relay Chat — which is clearly different in structure from Twitter's friend/follow network. Rangaswamy et al. [2010] initially attempted to categorize such groups into one of six categories: news, fun, cricket (sport), computing, educational, and business

[Rangaswamy et al., 2010]. Over time, such discussion categories begin to evolve; resulting in some groups having overlapping categories due to such evolution [Rangaswamy et al., 2010]. In the authors' words, some groups "while originating and evolving around a specific interest area, reach out to include various types of content from other groups as well, some even unrelated to the core interest area, to appeal to a wider audience and prospective members." [Rangaswamy et al., 2010]. They present a couple of case studies illustrating their point:

- **Sartaj group** (~200k members): started as a 'fun' group forwarding anecdotes and jokes, has evolved into using advertisements as a potential business model.
- **Vignesh Teacher** [*sic*] **group** (~65k members): started as an 'educational' group disseminating academic information, has evolved into being a prospective advertising tool for promoting jobs.

In closing, the authors concluded that "the entwining of 'fun' and 'business' [referring to the blurred boundaries of discussion group categories] further points to the seamless fusion of a variety of content categories" [Rangaswamy et al., 2010], which increases a group's potential for social networking.

Bentley and Metcalf [2009] have performed a field study on three kinds of sharing behavior exhibited by people in terms of mobile communication: motion (sharing one's location), music (sharing information about music currently listened to), and photos. By letting their test participants use customized Motorola phones to share these types of information, they infer that users directly or indirectly give away three types of cues about their current state:

1. **Motion presence**: gives hints to user location, activity, availability, destination (and estimated time of arrival).
2. **Music presence**: gives hints to user location, activity, and availability.
3. **Photo presence**: gives hints to user location, activity, and the presence of people around them.

These findings could be of benefit to microblogging and Twitter research, as Twitter provides (or emulates, with the help of third-party services) some of these 'presence sensors': e.g. location can be exposed by GPS data in tweets, and music/photos can be published using third party services such as *TwitPic* and *Last.fm*.

Battestini et al. [2010] performed a larger-scale study, in which the data set is comprised of ~58.2k SMS messages from 70 participants, aged between 17–26 years, over a five-month period. Battestini et al. [2010] first examined the type of contacts that form the bulk of SMS communication, and found out that all users send SMS messages to their friends; only ~60% of users to their family members, while only ~24% of users contact colleagues or workmates [Battestini et al., 2010].

Reasons for SMS chatter include asking questions or answering them, promoting 'ambient intimacy' with contacts, and chatter about everyday minutiae/trivium [Battestini et al., 2010]. Conversations are categorized in individual categories: with activity planning being the highest proportion (~31.7%), followed by relationships, chatting, school/jobs, places, and seeking information as the top six [Battestini et al., 2010]. However, simultaneous conversations were observed (i.e. an open-ended message sent to multiple recipients as a start), the categorization scheme starts to become different. The following is a non-ranked list of such categories proposed by Battestini et al. [2010]: future plans, sports scores, greetings, thank you messages, big incidents (e.g. robberies), looking for items, announcements, future communication (e.g. broadcasting new phone numbers), and chain letters (spam). Quantitative studies by Battestini et al. [2010] include tabulating average four-month SMS totals for a person ($\mu = 848.6$, $\sigma = 1000.5$); average message length ($\mu = 50.9$ characters, $\sigma = 46.2$); and average number of contacts per person (47.1 people, $\sigma = 35.3$).

Again, despite the differences in the network structure for SMS contacts and Twitter friends/followers, the types of messages sent and the comparative statistics for SMS messaging activity could be beneficial in related microblogging studies, justifying the literature survey covered within the subsection.

Adoption Habits

A paper on Facebook adoption by Bozkir et al [Bozkir et al., 2010] studied the average demographic of a real online social network — Facebook — complementing findings from Section 3.1.2 on Twitter's estimated demographics. Several statistics on a survey of 570 users are listed in Table 3.4.

Feature	**Statistics**
Gender	Ratio of males to females: roughly 1:1
Age bracket	74.1% in the (18–25) age bracket 20.53% in the (26–35) age bracket
Usage frequency	48.77% logs in at least once per day 25.26% logs in at least once per week
Daily usage	32.28% for 15 minutes or less on average 39.82% for 15–30 minutes
Educational background	70.35% are pursuing a Bachelor's degree 23.16% are in their Masters degrees
Group membership	99.82% part of a Facebook group (all but one user in their sample)

Table 3.4: Bozkir et al. [2010]: Demographic statistics on Facebook users, taken from a survey of 570 users.

The authors also hypothesized that several of these statistics have a degree of correlation with others. For example usage time has a correlation with gender/education level, while usage frequency correlates with age/period of daily use [Bozkir et al., 2010]. Bozkir et al. [2010] concluded that such association rules can be used to model adoption of Facebook as a social network.

3.5.2 Presence and Outward Sharing

This subsection discusses literature specifically dealing with microblogging usage intentions; in particular for promoting an online presence, and outward sharing of information.

Studies investigating primarily the message domain on Twitter have generally dealt with Twitter usage intentions and information-sharing. In their pioneering work on Twitter (Section 3.1.2), Java et al. [2009] proposed four kinds of Twitter usage intentions: "chatter, communication, information sharing, and news reporting".

Mischaud, in his Master's thesis [Mischaud, 2007], expanded the categorization with three more usage intentions: sending messages to contacts, publishing one's thoughts, and also share 'news-like information with others". This definition by Mischaud [2007] expanded on the original list by Java et al. [2009] by describing how Twitter is used to express ones' thoughts and describe what one is doing at a given moment. The categories of information sharing were also deconstructed into seven distinct groups [Mischaud, 2007]. Written in early 2007, the thesis [Mischaud, 2007] forecasted the current trend of using Twitter for more than just publishing statuses about everyday minutiae.

Naaman et al. [2010] have approached the study of microblogging usage intentions from a *social-awareness stream* (SAS) viewpoint. They characterize message content on Twitter (as a SAS) into categories, some of which have already been discussed by Java et al. [2009]; Mischaud [2007]. Notable concepts unique to Naaman et al. [2010] include "self-promotion, complaints, random thoughts, [posing] questions to followers, presence maintenance [and] anecdotes" [Naaman et al., 2010]. Here, Naaman et al. [2010] introduced the concept of using Twitter to maintain online presence and provide anecdotes to followers; which paved the way for future exploration of Twitter users' motivation in frequently posting tweets.

Several books on Twitter that focus on user participation and adoption of Twitter for marketing have also discussed about user intentions and online presence. Here, new concepts (apart from the ones in the previous paragraphs) will be given a summary. Comm [2009] wrote about the concept of 'mission accomplished' tweets to inform followers of accomplishments or milestones achieved (extending the findings on online presence [Naaman et al., 2010]), and picture distribution tweets (extending the concept of URL sharing [Java et al., 2009]). O'Reilly and Milstein [2009] discussed the need for "ambient intimacy" with friends and family as a result of presence maintenance on Twitter by answering the "what are you doing?" question. McFedries [2009] and Comm [2009] also highlight a current trend of Twitter usage, 'live tweeting', which is to tweet about events live as they unfold, e.g. conferences, trade shows and exhibitions.

A position paper by André et al. [2010] deals with the two topics of "well-being, and status feedback" in terms of microblog messages. In their preliminary study on status feedback using a Web interface which allows users to give ratings (negative, neutral or positive) to tweets by others, they found that the amount of positive feedback outnumber negative ones by a factor of approximately four. André et al. [2010] hypothesized that a Twitter user, in their words, "only follow people they are interested in... are friends with many of the people they follow and thus are likely to 'play nice', and people are more comfortable giving positive feedback"[André et al., 2010]. In another experiment, André et al. [2010] studied the notion of *wellbeing* with tweets using a custom tagging system. From this experiment, participants reported "anecdotes of value in self-reflection at the time of update... [and also after]" [André et al., 2010], connoting the shift of personal state and experience online by users engaging in microblogging.

Subramanian and March [2010] have performed two studies "to understand what people want to share, with whom, and what challenges they currently face with existing sharing applications" [Subramanian and March, 2010]. The first dealt with the researchers 'shadowing' ten test subjects in real-life for a few hours each, capturing photographs and recorded their activities with the subjects' full knowledge. The second study involves nine participants utilizing an iPhone-based photo-sharing application which allows them to take pictures (15–30 daily), and annotate them with a description of what to share and to whom. From these studies, they have found that:

- Twitter usage has evolved from originally "sharing day-to-day intimate details with a small close-knit group... to sharing the mundane and intimate with both close friends and complete strangers" [Subramanian and March, 2010];
- status updates on Twitter are often "carefully crafted" to reflect different aspects or personas of oneself;
- sharing habits are also influenced by different motivations, individual styles, and target audiences of their tweets;
- current microblog technologies cannot target specific groups of people for different types of updates, as opposed to social networks with custom privacy controls, such as Facebook
- users have difficulty in defining clear groups from their contacts in the first place;
- users tend to 'push' only interesting content to strangers while allowing people close to them to 'pull' more personal or trivial content; and
- it is difficult to filter the level of detail to their different target audiences (e.g. friends versus work superiors).

Ramsden [2008] conducted a survey of 17 people describing "how and why [they use] Twitter", and revealed three main motivations for using a microblog service such as Twitter: to keep other people updated with their goings-on (~60% of their sample), directed communications with other users (using public `@user` messages, ~22% of their sample), and to publish statuses for personal consumption (e.g. observations/ideas, ~7% of their sample). Twitter is also sometimes used in conjunction with (or integrated into) other services, such as Facebook. An interesting finding relating to Subramanian and March [2010] above is that ~82% of the surveyed users are conscious of "of the different audiences that followed them" [Ramsden, 2008] and try to avoid publishing e.g. overtly personal statuses for fear of them reaching the wrong audience.

3.5.3 Privacy Concerns

With the high instances of outward sharing on Twitter, user privacy remains an issue which necessitates attention to, as with all types of communication and social networking. Such an issue has been explored in academic literature, and the consequences of privacy breaches are commonplace in mass media.

Lawler and Molluzzo [2011], who authored a study on first-year college student perceptions of privacy on Facebook, MySpace and Twitter, revealed that on the whole, privacy on online social networks are not clearly known nor understood amongst its users. Despite their findings mainly focusing on personally-identifiable information on Facebook, certain pieces of information (such as real name and current location) are similarly available from a microblogging site such as Twitter.

Humphreys et al. [2010] highlighted the fact that combinations of certain information on Twitter have unexpectedly caused a privacy breach resulting in negative side effects: a burglary case in 2009 comes to mind, where the burglar allegedly knew his victim was away thanks to the victim's Twitter updates. Humphreys et al. [2010] are of the opinion that users of services such as Twitter fail to inform themselves that privacy settings can be changed and that changes to privacy settings may be deemed too difficult. Humphreys et al. [2010] highlighted that in August 2009, fewer than 8% of the total users have private accounts, a sheer contrast to the time period of January 2007 where ∼40% of Twitter users are private.

As per Lawler and Molluzzo [2011], Humphreys et al. [2010] draw attention to the fact that temporal information and real names are readily obtained from a public user account. Their experiment focused on manually coding 728 tweets for presence of specific personal details. Most of their dataset did not contain any personally identifiable information; however Humphreys et al. [2010] found out that co-occurrences of certain features that can ascertain someone's presence at a given time or place, such as:

- **Self-references**: personal pronouns (e.g. my, I)
- **Personally-identifiable information**: such as contact details
- **Names**: including both nicknames and real names
- **Locations**: found in ∼10% of their sample in conjunction with personal pronouns
- **Time**: specific nouns or date ranges were found in ∼15% of their sample in conjunction with personal pronouns

A startling conclusion drawn from Humphreys et al. [2010] is that ∼3% of overall tweets contains mentions of all three features: self-references, location, and time. In the worst-case scenario, the ∼3% figure extrapolates to about ∼360k potential privacy-risking tweets daily, and if data from a user's profile is used together with the tweet, the risk of privacy breaches may be compounded [Humphreys et al., 2010].

‘Zooming’ and Selective Broadcasting

An issue identified in Subramanian and March [2010], Ramsden [2008], and McAllister [2010] is the inability of Twitter and current microblogging services to allow selective filtering of information. As described above, for example, certain users might want to send detailed private tweets for friends and family; choosing to provide a more vague description for workplace colleagues; and blocking it for everyone else.

Subramanian and March [2010] wrote that such selective filtering affords the “benefit of plausible deniability and allows for the notion of ‘saving face’... and allows the right level of information to be delivered to each audience member” [Subramanian and March, 2010]. As a proposed solution, existing issues can be alleviated by the creation of a ‘zoom’ feature that allows the user to customize the level of ‘zoom’ (detail) for each group of contacts they have [Subramanian and March, 2010].

3.5.4 Social Information Needs and Wants

Two papers reviewed in this section — Kaufman and Chen [2010]; Wilson [2008] — investigated how people utilized Twitter as an information source; in other words, conducting ‘social search’ on Twitter. ‘Social search’ mentioned earlier refers to a person’s intrinsic need for information from a social context; as opposed to user search (Section 3.4.4) referring to a particular user’s search experience on Twitter.

Kaufman and Chen [2010] investigated the role of Twitter “as a tool for capturing comprehensive locality information” [Kaufman and Chen, 2010]. This role was fulfilled by Kaufman and Chen [2010], with a proposal to use Twitter for provision of location-specific information, so that a newcomer (to the area) can better understand his/her surroundings. The proposal, which revolved around the usage of a mobile client, has two modes of operaiton. “*Explore City*” is used to reveal places of interest, food, and information about a city based on localized tweets; arguably better than conventional geographic tools such as Google Maps [Kaufman and Chen, 2010]. “*Explore Space*” on the other hand focuses on one particular venue of interest, allowing the user to explore detailed information from others’ tweets about that venue [Kaufman and Chen, 2010]. Although still in the phases of a conceptual design, this research has shown the effectiveness of Twitter in providing social location-based information.

Wilson [2008] performed an early “analysis into how people describe and converse about their own information needs” on Twitter. By analyzing about ~189k unique tweets from about ~163.5k users (with approximately ~15k being retweets), Wilson [2008] checked for occurrences and frequency of “...different terms as people describe their searching actions”, and how the different terms relate to different search interests [Wilson, 2008]. By order of popularity, the top six keywords/terms describing search and types of information sought are as follows:

1. **studying**: exams, scientific studies;

2. **hunting**: sport;

3. **finding**: ‘finding out’ about things;

4. **searching**: food, people, music/pictures, friends;

5. **looking**: people, jobs, technology; and

6. **exploring**: places.

From their analysis, Wilson [2008] provided an insight on the types of information sought on Twitter, as well as the different synonyms frequently associated with social information needs on Twitter.

Banerjee et al. [2009] have presented a novel approach to discovering user interest and context by analyzing contents of tweets (in the message domain) from the Twitter Search API. The user domain plays a role to isolate messages from users who are active (based on their frequency of writing tweets) and among specific major cities in the western hemisphere. Their objective is to obtain “tweets that capture a user’s real-time interests in activities” [Banerjee et al., 2009] by searching for three types of words: activities (e.g. dance, food, movie, music, sports), action verbs (e.g. watch, play), and temporal nouns (e.g. today, tonight). By matching co-occurrences of such keywords in tweets, Banerjee et al. [2009] managed to observe user interests and planned activities in different cities, proving that Twitter is suitable for user context analysis in terms of user interest, emotions, presence, etc. with the objective of capturing consumer data in real-time.

3.6 Practical Applications and Usage of Twitter

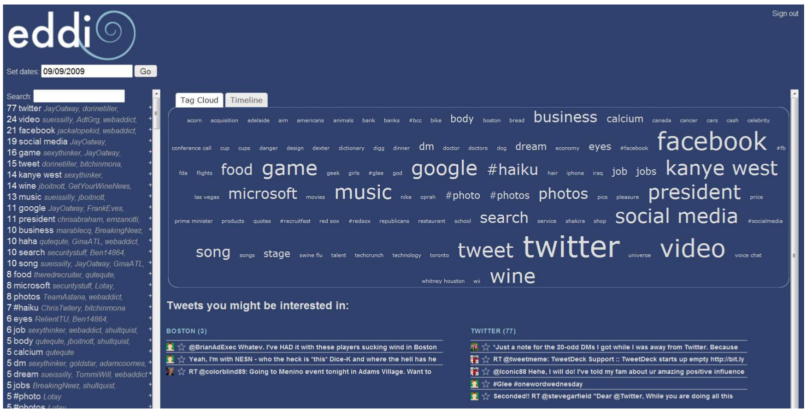
The sixth and final theme of literature surveyed in Chapter 3 is on practical applications and usage of Twitter. This will provide one with an overview of how Twitter can be used ‘in the real world’ by humans and organizations alike; how Twitter data can be consumed effectively from the perspective of computer-human interaction; and emerging fields of future practical Twitter research.

3.6.1 Visualization and Computer-Human Interaction (CHI)

Applications of Twitter data from both users and messages in terms of visualization, CHI, and electronic art are commonplace. Research ideas in this regard normally take both user and message domains into account, where information is presented in a relevant and easy to understand format, e.g. a visualization of tweets based on specified criteria, or to study a user based on his individual tweets.

Research in terms of CHI with respect to the visualization of tweets, the subject reviewed within this section, has a focus on making user interaction with Twitter intuitive, graphical, and easy-to-understand. From the earlier discussion of Bernstein et al. [2010] in Section 3.4.2 — cf. the amount of tweets received by an active Twitter user is too much for daily consumption [Bernstein et al., 2010] — I shall now move on to an elaboration of the need for graphical visualization of Twitter feeds.

The proposed *Eddi* system by Bernstein et al. [2010], which was developed based on user surveys, allows users to browse tweets aggregated by topics, personalized based on their criteria for value: e.g. relevance to interests, connection to other users, or *serendipity* [Bernstein et al., 2010]. Figure 3.1 shows an example of the *Eddi* visualization prototype.



77 twitter JayOatway, dorinetiller,
24 video sueissilly, AdtGrg, webaddict,
21 facebook jackalopekid, webaddict,
19 social media JayOatway,
16 game sexythinker, JayOatway,
15 tweet dorinetiller, bitchinmona,
14 kanye west sexythinker,
14 wine jboitnott, GetYourWineNews,
13 music sueissilly, jboitnott,
11 google JayOatway, FrankEves,
11 president chrisabraham, emzanotti,
10 business marablecq, BreakingNewz,
10 haha qutequte, GinaATL, webaddict,
10 search securitystuff, Ben14864,
10 song sueissilly, JayOatway, GinaATL,
8 food theredrecruiter, qutequte,
8 microsoft securitystuff, Lotay,
8 photos TeamAstana, webaddict,
7 #haiku ChrisTwitery, bitchinmona
6 eyes RelientTU, Ben14864,
6 job sexythinker, webaddict, shultquist,
5 body qutequte, jboitnott, shultquist,
5 calcium qutequte
5 dm sexythinker, goldstar, adamcoomes,
5 dream sueissilly, TommiWill, webaddict
5 jobs BreakingNewz, shultquist,
5 #photo Lotay
5 #photos Lotay

acorn acquisition adelaide aim americans animals bank banks #bcc bike body boston bread business calcium canada cancer cars cash celebrity conference call cup cups danger design dexter dictionary digg dinner dm doctor doctors dog dream economy eyes #facebook facebook #fb fda flights food game geek girls #glee god google #haiku hair iphone iraq job jobs kanye west las vegas microsoft movies music nike oprah #photo #photos photos pics pleasure president price prime minister products quotes #recruitfest red sox #redsox republicans restaurant school search service shakira shop social media #socialmedia song songs stage swine flu talent techcrunch technology toronto tweet twitter universe video voice chat whitney houston wii wine

Tweets you might be interested in:

BOSTON (3)

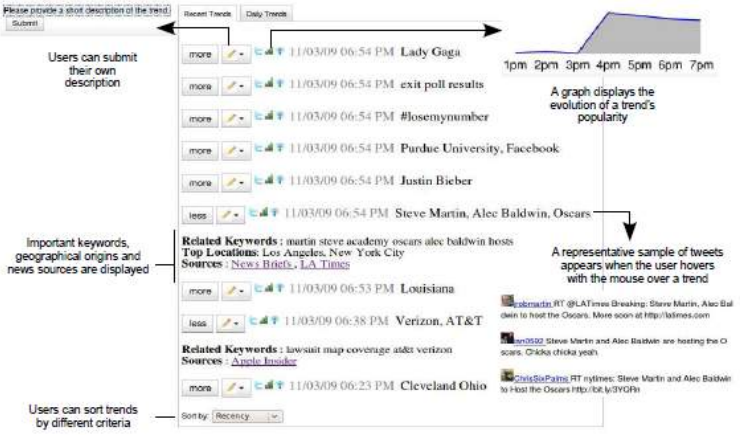
☆ @BrianAdExec Whatev. I've HAD it with these players sucking wind in Boston
☆ Yeah, I'm with NESN - who the heck is *this* Dice-K and where the hell has he
☆ RT @colorblind89: Going to Menino event tonight in Adams Village. Want to

TWITTER (77)

☆ *Just a note for the 20-odd DMs I got while I was away from Twitter. Because
☆ RT @tweetmeme: TweetDeck Support :: TweetDeck starts up empty http://bit.ly
☆ @Iconic88 Hehe, I will do! I've told my fam about ur amazing positive influence
☆ #Glee #onewordwednesday
☆ Seconded!! RT @stevegarfield "Dear @Twitter, While you are doing all this

Figure 3.1: Bernstein et al. [2010]: Personalized *Eddi* Twitter interface, displaying a tag cloud containing personalized topics of interest for a user.

The paper by Mathioudakis and Koudas [2010] demonstrated *TwitterMonitor*, another Twitter visualization tool. Compared to *Eddi*, *TwitterMonitor* does not personalize streams per user; rather it reflects the overall trend of the current public timeline of messages. It attempts to group 'bursty' keywords into related groups based on their co-occurrences within the live Twitter message stream, parsing about 10 million messages per day. Such 'bursty' keyword groups are then analyzed using content extraction algorithms to label them with appropriate keywords that reflect the overall trend by the group. Their live demonstration version at the *SIGMOD'10* conference is illustrated in Figure 3.2.


Please provide a short description of the trend.
Submit
Recent Trends
Daily Trends
Users can submit their own description
1pm
2pm
3pm
4pm
5pm
6pm
7pm
A graph displays the evolution of a trend's popularity
Important keywords, geographical origins and news sources are displayed
A representative sample of tweets appears when the user hovers with the mouse over a trend
Users can sort trends by different criteria
Sort by: Recency

more 11/03/09 06:54 PM Lady Gaga
more 11/03/09 06:54 PM exit poll results
more 11/03/09 06:54 PM #losemynumber
more 11/03/09 06:54 PM Purdue University, Facebook
more 11/03/09 06:54 PM Justin Bieber
less 11/03/09 06:54 PM Steve Martin, Alec Baldwin, Oscars

Related Keywords : martin steve academy oscars alec baldwin hosts
Top Locations: Los Angeles, New York City
Sources : News Briefs , LA Times

more 11/03/09 06:53 PM Louisiana
less 11/03/09 06:38 PM Verizon, AT&T

Related Keywords : lawsuit map coverage at&t verizon
Sources : Apple Insider

more 11/03/09 06:23 PM Cleveland Ohio

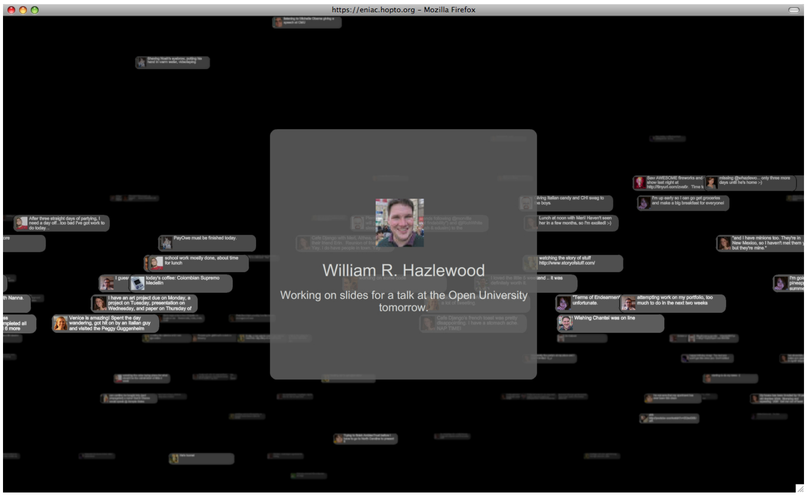
robmartin RT @LATimes Breaking: Steve Martin, Alec Baldwin to host the Oscars. More soon at http://latimes.com

ian0502 Steve Martin and Alec Baldwin are hosting the Oscars. Chicka chicka yeah

ChrisSixPalms RT nytimes: Steve Martin and Alec Baldwin to Host the Oscars http://bit.ly/3YQRn

Figure 3.2: Mathioudakis and Koudas [2010]: *TwitterMonitor* visualization, showing groups of tweets, collated by keyword, in real-time.

Similar to *TwitterMonitor*, *TwitterSpace* by Hazlewood et al. [2008] is another "public display of tweets" [Hazlewood et al., 2008]. What makes it unique is that *TwitterSpace* is a visualization of tweets "published by members of [their] local community" [Hazlewood et al., 2008]. Recent tweets by followers of a particular Twitter account (set up for the purposes of the project) are set up in a timeline-based interface to visualize the chatter of users who ‘belong’ to this ‘community’ as seen in Figure 3.3. This creates a “community-at-a-glance... [which aims to blend] the virtual space of Twitter with [their] physical community centers” [Hazlewood et al., 2008].


https://eniac.hopto.org - Mozilla Firefox

William R. Hazlewood

Working on slides for a talk at the Open University tomorrow.

After three straight days of partying, I need a day off...too bad I've got work to do today...

PayOwe must be finished today.

school work mostly done, about time for lunch

today's coffee: Colombian Supremo Medellin

I have an art project due on Monday, a project on Tuesday, presentation on Wednesday, and paper on Thursday of

Venice is amazing! Spent the day wandering, got hit on by an Italian guy and visited the Peggy Guggenheim

Saw AWESOME fireworks and show last night at http://tinyurl.com/zva6r. Time t

missing @whazlewo... only three more days until he's home :-)

Giving Italian candy and CHI swag to the boys.

I'm up early so I can go get groceries and make a big breakfast for everyone!

Lunch at noon with Meri! Haven't seen her in a few months, so I'm excited! :-)

"and I have minions too. They're in New Mexico, so I haven't met them but they're mine."

watching the story of stuff http://www.storyofstuff.com/

I loved the little 5 weekend .. it was definitely worth it.

"Terms of Endearment" unfortunate.

attempting work on my portfolio, too much to do in the next two weeks

Wishing Chantel was on line

Figure 3.3: Hazlewood et al. [2008]: *TwitterSpace* serves to be a public timeline of ‘local community’ tweets. This example shows a timeline of tweets by people participating in the *TwitterSpace* project.

Lastly, a noteworthy collection of visualization algorithms by Donath et al. [2010], the first two of which are directly related to Twitter, will be discussed. These five algorithms [Donath et al., 2010] are noteworthy in that they visualize social media data (Twitter inclusive) in a novel and artistic fashion:

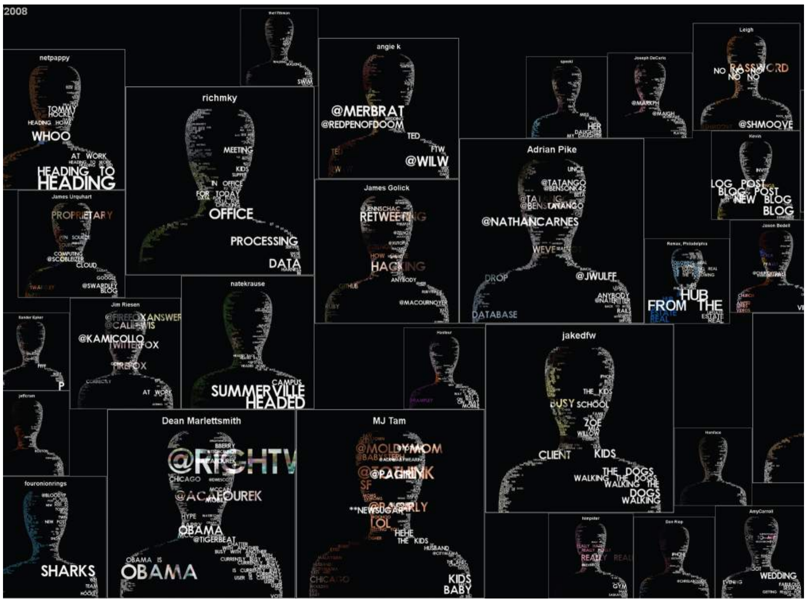


Figure 3.4: Donath et al. [2010]: *Lexigraphs* illustrating frequently-occurring keywords by users, designed to look like portraits.

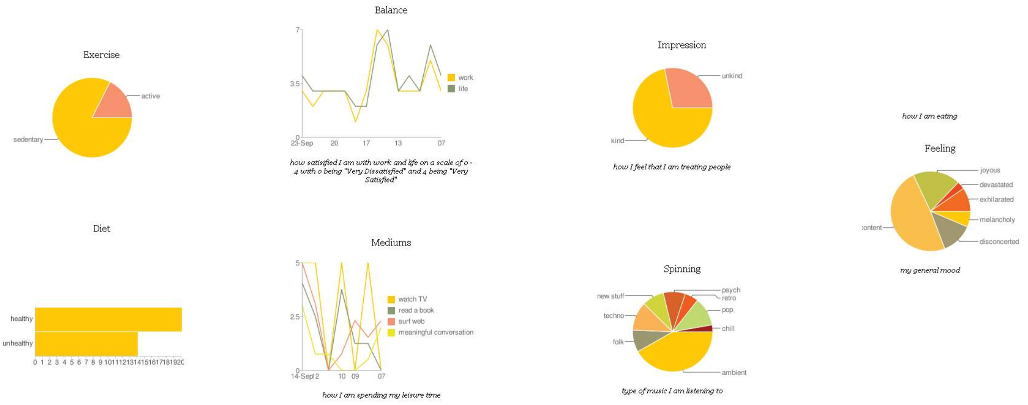


Figure 3.5: Donath et al. [2010]: *Mycrocosm* visualizing a user's personal statistics (from their tweets), in the form of annotated graphs.

- ***Lexigraphs*** (Figure 3.4): a "group portrait of users of Twitter... shown as a silhouette outlined in words derived from their updates [and] animated by the rhythm of their postings" [Donath et al., 2010]. This provides the viewer with an opportunity to view the Twitter user from the perspective of his conversation topics on Twitter; the researchers however choose to populate the portrait with "words the subjects use with unusual frequency" [Donath et al., 2010]. *Lexigraphs* is shown in Figure 3.4.

- ***Mycrocosm*** (Figure 3.4): an introspective visualization of a user's own "everyday 'personal statistics' using simple graphs to display their data" [Donath et al., 2010] where users can freely choose what kind of information is exhibited on Twitter, and by extension, for *Mycrocosm*. *Mycrocosm* is shown in Figure 3.5.

- ***AuthorLines* and *Themail***: illustrations of the timeline of reply and communication habits of users in a discussion forum and in email, respectively. These 'data-portraits' enable a user to understand his or her own behavior, and allows others in the community to also discern an individuals' behavior in the context of the community.

- ***Conversation Maps***: akin to conventional tag clouds that highlight important words in a users' textual communications, with the added twist of having these tag clouds linked to other tag clouds representing other users; in other words it can be seen as a network of tag clouds pertaining to a conversation a user has with his contacts.

- ***PeopleGarden***: employs word-detection algorithms to determine the emergent features of messages from a particular community (e.g. emotion or political affiliation) to visualize its participants in terms of a 'garden'.

I opine that the latter three visualizations can be adapted in future research to deal with microblog data, as Twitter allows the easy exploration of communities (e.g. the *Lists* feature, or other forms of community detection cf. [Java et al., 2009]). The conversations and chatter between users also are easily available from the user social graph, or from related research e.g. [Ritter et al., 2010].

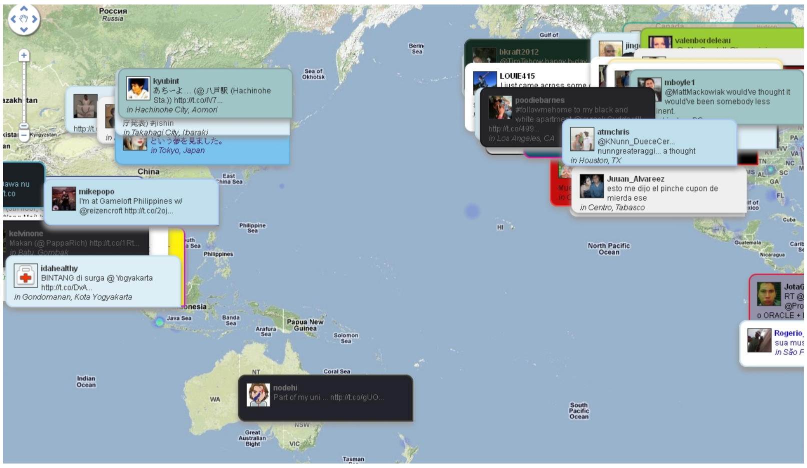
Commercial Website Visualizations of Twitter Data

Besides academic work on Twitter-based visualizations, there are several interactive Web 2.0 sites or mashups that attempt to make sense of the large volume of information on Twitter. This subsection of the literature review will detail several notable ones.

1. *TwitterVision* [Troy, 2011] (Figure 3.6) is a mash-up between the Twitter public timeline and the Google Maps API to visualize Twitter updates based on the published geographic location and superimposes this onto a Google Map display to have a real-time display of Tweets based on their geographic location. The user location is obtained via a set of coordinates generated by GPS-enabled devices or browsers.

2. Bloch and Carter [2009] from the New York Times published an experimental Flash applet — a geographically-distributed *tag cloud* of sorts — that visualizes Twitter activity during the 2009 Super Bowl. This is accomplished by mapping out the location and frequency of commonly used words in Super Bowl related messages on a map of the United States (Figure 3.7). This is not dissimilar with the use of geography and time to track the spread of a current real-life event (Section 3.2).

3. The concept of timeline visualization has also been implemented using keywords, hashtags, and trending keywords in Twitter. This is illustrated by the use of timelines in websites such as *TwitScoop* [Lollicode SARL, 2009] and *What The Trend?* [Mayer, 2009]: interactive timelines are coupled with other elements (such as a tag cloud, or list of related tweets) to highlight the prevalence of popular words in current Twitter activity.



kyubint
あちーよ… (@八戸駅 (Hachinohe Sta.)) http://t.co/lV7...
in Hachinohe City, Aomori

庁発表) #jishin
in Takahagi City, Ibaraki

という夢を見ました。
in Tokyo, Japan

mikepopo
I'm at Gameloft Philippines w/ @reizencroft http://t.co/2oj...

kelvinone
Makan (@ PappaRich) http://t.co/1Rt...
in Batu, Gombak

idahealthy
BINTANG di surga @ Yogyakarta http://t.co/DvA...
in Gondomanan, Kota Yogyakarta

nodehi
Part of my uni ... http://t.co/gUO...

bkraft2012
@TimTebow happy b-day

LOUIE415
I just came across some

valenbordeleau

poodiebarnes
#followmehome to my black and white apartment
http://t.co/499...
in Los Angeles, CA

mboyle1
@MattMackowiak would've thought it would've been somebody less

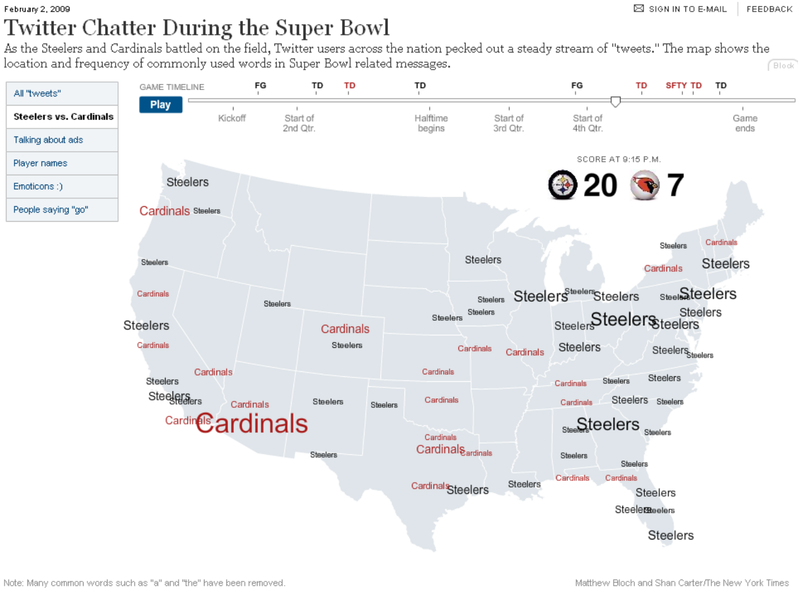
atmchris
@KNunn_DueceCer... nunngreateraggi... a thought
in Houston, TX

Juuan_Alvareez
esto me dijo el pinche cupon de mierda ese
in Centro, Tabasco

JotaG
RT @
@Pro
o ORACLE +

Rogerio_
sua mus
in São P

Figure 3.6: [Troy, 2011]: *TwitterVision*, showing 'live' tweets overlaid on a world map.



February 2, 2009

Twitter Chatter During the Super Bowl

As the Steelers and Cardinals battled on the field, Twitter users across the nation pecked out a steady stream of "tweets." The map shows the location and frequency of commonly used words in Super Bowl related messages.

Note: Many common words such as "a" and "the" have been removed.

Matthew Bloch and Shan Carter/The New York Times

Figure 3.7: [Bloch and Carter, 2009]: New York Times' 2009 Super Bowl visualization tool, using spatio-temporal perspectives to visualize Twitter chatter.

3.6.2 Microblogging in Organizations

Microblogging started as a form of intra-organization communication, as Twitter was developed as an internal communication tool at Odeo [O'Reilly and Milstein, 2009]. Due to its humble beginnings, the rapid expansion and popularity of Twitter has necessitated recent research to study the characteristics of microblogging in an organizational context, where it all began.

Thom-Santelli et al. [2010] studied the IBM *BeeHive* internal lightweight microblogging network from the cross-cultural perspective of three IBM branches: China, India and the United States. The study comprised approximately ∼60k users (between ∼6k to ∼13k active users at a given time) with a total of ∼150k comments posted. They found out that the microblogging behavior among users differ based on their cultural norms, for example the Indian branch of IBM has users which post informal or more expressive posts; as compared to the ones in the US site. They conclude that "familiarity with the characteristics of other social networking/microblogging sites" [Thom-Santelli et al., 2010] influences the internal organizational microblogging behavior of users. For example, US users are more familiar with Twitter tend to post updates on 'what are you doing?' compared to Indian users' adoption of Orkut (a similar microblogging service) which have more personally-expressive posts. Real-world cultural *power distance* also plays a role in influencing the types of status messages created [Thom-Santelli et al., 2010].

In the context of companies just starting to adopt microblogging, Zhang et al. [2010] performed a survey on the adoption of microblogging using the *Yammer* internal microblogging service. By studying a *Fortune 500* company by means of a 13-month data log on *Yammer* and some interviews with adopters, they found out that adoption of microblogging in an organization grows through four progressive stages, i.e. "initial adoption (registering an account), continued use (logins for either reading or posting), contributing (posting content/*following* others), [and finally] promoting (inviting others)" [Zhang et al., 2010]. They also found out that hubs in the user network — "individuals who invite many people to participate" [Zhang et al., 2010] — play a vital role in the adoption of microblogging. Further study is still necessary to investigate whether these hubs correspond to superiors or high-ranking employees.

Ehrlich and Shami [2010] performed a comparative survey to distinguish between internal organizational microblogs and Twitter by analyzing the contents of over ~5k microblog messages by employees of an organization, which are randomly sampled. This constituted approximately ~3.1k messages from Twitter and ~2.2k originating from an internal microblog site; these were authored by 1257 users, of whom 25 are interviewed. Prior research [Java et al., 2009; Zhao and Rosson, 2009] has come up with four categories of chatter, conversation, sharing (of information), and news; which are then adapted by Ehrlich and Shami [2010] to create a new list of six categories:

- Status (i.e. answering 'what are you doing?');
- Information (comments, opinions, news, and links);
- Retweets;
- Asking questions;
- Directed tweets (`@user` messages); and
- Directed questions (combination of both directed tweets and questions);

The tweets were manually coded to determine which category suits each individual one; and the time and presence of internal organizational jargon or lingo are studied to provide context for tweets. The results of their study is as follows [Ehrlich and Shami, 2010]:

- **Internal/organizational microblogs**: used to chat about company-related subjects, things about work, to get to know colleagues, help colleagues solve problems, and to connect with colleagues (in the context of mobile workers). This form of communication can be summarized as providing information and enabling phatic communications (with less 'noise' and background chatter)
- **Public microblogs such as Twitter**: used to discuss news in real-time, to have conversations/chats with other people, publish personal statuses, directed conversations, and mentioning trivium and information.

3.6.3 Applied Microblogging in Science, Education, and Governance

This subsection briefly details the role of applied microblogging in science, education (especially the tertiary sector), and governance; as descriptions of such applications in current literature provides a better understanding of current microblogging practice by such institutions.

Science

Vertesi [2010] wrote a position paper on the usage of microblogging by NASA in building a public presence. Examples of Twitter use by NASA include the creation of Twitter accounts for each of their Mars Rovers as their online 'personas'. Several interaction patterns have been discussed by Vertesi [2010] in this regard, such as: differentiation of private versus organizational tone of voice for such Twitter accounts, the behavioral patterns of retweets from these accounts; their dissemination of information (retweets and URLs); and the scale of followers for such accounts (tens of thousands, with the highest being about ~40k for the *Phoenix Rover*).

Government

Wigand [2010] authored a paper on the adoption of Twitter by government agencies in the United States. This paper starts with statistics of current Twitter adoption in the USA: almost 20% of Twitter users are from the USA. Based on *GovTwit*, a directory of governmental 'users' of Twitter, 2,349 users have contributed to more than ~192k tweets with about ~28 million followers in total [Wigand, 2010]. Several examples of US government agencies who have adopted Twitter to reach out to citizens are NASA's *Phoenix Rover* account (cf. Vertesi [2010]), the Armed Forces Personnel Administration Agency, and the US State Department ("one of the main channels to disseminate information about the [Haiti earthquake] emergency" [Wigand, 2010]). According to Wigand [2010], four major roles of Twitter use in US governmental agencies have been identified as: (1) extending the reach of communication; (2) the updating, broadcasting and sharing of information; (3) the building of relationships; and (4) for "collaborating with stakeholders" [Wigand, 2010].

Education

Du et al. [2010] opined that Twitter in the classroom empowers each individual student with the 'right' to say something and express themselves. Ebner et al. [2010] in a case study on microblogging use in a tertiary setting, found that an average of 7.5 posts per student per working day were generated; of which communication between students and teachers forms a high percentage (~60%), and that content which dealt with course administration formed ~19% of the total. Dunlap and Lowenthal [2009] highlighted several use cases of microblogging in an educational environment, such as:

- asking questions to peers, educators, the community, and even experts (cf. Ramsden [2008]);
- facilitating communication (cf. Ebner et al. [2010]);
- promoting information sharing, commenting, and dissemination; and
- 'conference blogging' or broadcasting live updates from an academic conference (cf. 'live tweeting' Comm [2009] in Section 3.5.2)

The preceding subsection provides a summary overview of the increasing adoption of microblogging in educational settings. There are many more pieces of published research containing detailed analyses of Twitter as a facilitator for learning; these are not covered in this chapter as it is beyond the scope of my literature review.

3.6.4 Potential Fields of Emerging Microblogging Research

Before I conclude this chapter, I note several potential fields of emerging research related to microblogging that have recently been suggested [Böhringer and Gluchowski, 2009]. Such research topics are predicted to be promising areas of future research:

- **text mining and semantic analysis on short microblog messages**: the short 140-character length of microblog messages complicates traditional forms of text mining

- **complex event processing**: "each individual micro-blogging posting is in itself constitutes an event" [Böhringer and Gluchowski, 2009] and therefore would be suitable as input for potential complex event processing

- **architecture decentralization and security**: research on this has already started. Xu and Farkas [2008] identified centrality as the weakness to a microblogging service, due to: threats to stability such as from denial-of-service attacks, bottlenecks on the system's performance resulting in measures such as 'rate limiting' (which impacts the amount of data that can be collected for research), and single points of failure which can cause the whole microblogging service to fail. Xu and Farkas [2008] have designed a working prototype of decentralized microblogging service as a proof-of-concept as a result of their findings.

3.7 Concluding Notes on State of the Art

Throughout this chapter, I have surveyed new research on Twitter and microblogging from circa 2009 till early 2013. Since then, there is a mushrooming of literature and related research on the subject, as well as applications that are hitherto not considered for research from the perspective of microblogging and social media. I have, in this chapter, also identified several emerging topics of theoretical research (e.g. decentralization of the microblogging architecture), and also up-and-coming practical applications of Twitter (e.g. in government, education, activism, and for promoting democracy).

Also, this chapter has explored the idea of two separate (yet interdependent) domains of the *user* and the *message* in microblogging services [Cheong and Lee, 2010a; Cheong and Ray, 2011; Cormode et al., 2010]. I have also covered the significance of both these domains, their relation to one another, and their interdependence in our evaluation of current literature. Research that merges the study of both these domains are still lacking; in spite of that, several promising research studies reviewed in this chapter have leveraged the combination both domains in pattern detection and classification.

In terms of all six themes surveyed in this chapter, my contributions in the rest of this thesis will serve to fill in gaps and contribute to the body of knowledge. These, in order of discussion in this chapter, are: more extensive exploratory studies on Twitter (Chapter 5); a better understanding on the spread of information on Twitter (Chapter 7); applications of pattern recognition for the revelation of emergent behavior (Chapter 6); modeling and detection of sentiment with regard to trends on Twitter (Chapter 8); human factors on Twitter (Chapter 4); and practical applications of Twitter (Chapters 4, 5, and 7).

Chapter 4

Uncovering Inferences from Twitter Metadata

"There are times when all the world's asleep,
The questions run too deep, for such a simple man,
Won't you please, please tell me what we've learned?"

— Supertramp,
The Logical Song (1979).

Parts of this chapter have been published as:

Cheong, M. and Lee, V. [2009]. Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base, Proc. CIKM 2009 Co-Located Work- shops: SWSM 2009, pp. 1–8.

Cheong, M. and Lee, V. [2010c]. Twitmographics: Learning the Emergent Properties of the Twitter Community, *From Sociology to Computing in Social Networks: Theory, Foundations and Applications, Vol. 1 of Lecture Notes in Social Networks*, Springer-Verlag, pp. 323–342.

Cheong, M. and Lee, V. [2011]. A Microblogging-based Approach to Terrorism Informatics: Exploration and Chronicling Civilian Sentiment and Response to Terrorism Events via Twitter, Information Systems Frontiers **13**(1): 45–59.

Cheong, M., Ray, S. and Green, D. [2012b]. Large-scale Socio-demographic Pattern Discovery on Microblog Metadata, *Proc. SoCPAR 2012.*

In the previous chapter, I have surveyed the extant literature to identify the state-of-the art in research, with respect to Twitter. Existing studies specialize only on either the user or the message domain, but rarely both. There is a dearth of research dealing with the combination of both domains, with emphasis on the analysis of the raw metadata themselves, and methods in which such raw data can be transformed into useful heuristics and information.

In this chapter, my goal is to examine the inner workings of Twitter to discover how both overt and latent metadata on Twitter can be used as a data source for mining socio-demographic inferences, and also for detection of emergent patterns among the tweets generated by its user base. This contributes to the solving of **Subgoal 2** of my overall thesis.

Firstly, in Sections 4.1 and 4.2, I am going to discuss the two Twitter Application Programming Interfaces (APIs): the older on-demand APIs, and the newer *Streaming API*. Based on both existing literature and my original research, I will describe how the different APIs on Twitter can be used in tandem with one another for programmers and researchers to access the vast amount of metadata on Twitter. I will then describe several issues pertaining to the suitability of the APIs for my research, including weaknesses and workarounds.

Next, in Section 4.3, I will describe the complete set of metadata (properties) from the central Twitter domains — the user domain and message domain — that is made readily available by the Twitter APIs. This includes a coverage of the metadata format, as well as methods for using them. Section 4.4 then describes my preliminary investigation into the interdependence of the two domains, and how these domains are supported by Twitter APIs.

This leads into Section 4.5, where I introduce a bespoke dataset I created — the *10-Gigabyte Dataset*— which is used throughout this thesis (specifically the latter half of the current chapter) as sample data for the discovery, testing, and validation of inference algorithms and metrics.

The heart of this chapter lies in Sections 4.6, 4.7, and 4.8. These sections detail my algorithms and metrics for transforming the raw metadata from both domains into valuable statistics or inferences based on my research and empirical observations. My contributions with regard to this can be divided into three main areas: real-life demographic properties (Section 4.6); online presence of users (4.7); and tweeting/communication patterns (Section 4.8).

4.1 Twitter's On-demand APIs

For data mining to be conducted on Twitter metadata, one must first need to understand how Twitter exposes its metadata to its end users or researchers. Therefore, this section contains an overview of the core Twitter Application Programming Interfaces (APIs), their properties, and how one can make use of such APIs for metadata retrieval.

4.1.1 Overview of On-demand APIs (currently deprecated)

In circa 2009 during my initial research for this thesis, the Twitter API was split into two main APIs, each closely tied to one of two central domains in Twitter [Krishnamurthy, 2009]. Both of these are REST (*REpresentational State Transfer*) APIs. In a nutshell, REST refers to the fact that these Twitter APIs process the request for user information and returns it in a *representational* manner, or in programming terms, a metadata 'object'.

These two core REST APIs (of which other functions or sub-APIs are derived from) are:

1. **The REST-`user` API**: Twitter's originally-developed REST API which exposes the *user* object; it is accessed via the `users/show` method.

2. **The `search` API**: This API, which Twitter acquired from Summize Inc. [Twitter Inc., 2012a], allows users to search for tweets and also to identify Twitter Trends. The `search` method retrieves *messages* based on search criteria provided by the user[1].

There are, however, a few subtle differences between the REST-`user` API and the `search` API, though.

As described by Twitter Inc. [2011a],

> "*[The API differences are] entirely due to history... [as] Summize, Inc. was originally an independent company that provided search capability for Twitter data. Summize was later acquired and rebranded as Twitter Search. Rebranding the site was easy, [however] fully integrating Twitter Search and its API into the Twitter codebase is more difficult. It is in our pipeline to unify the APIs, but until resources allow the REST-`user` API and Search API will remain as separate entities.*"

4.1.2 Interdependence of the Two APIs on Twitter

During the early stages of this PhD research [Cheong and Lee, 2010c], I initially mapped out several relationships between the two separate domains on Twitter, and their relation to the above two APIs. Figure 4.1 is an annotated UML interaction diagram that illustrates the interdependencies of the two domains, represented here by two main APIs.

The *Trends* API mentioned in Figure 4.1 is a simple API in the message domain (which constitutes a part of the `search` API), that reveals the top ten *Trending Topics* discussed on Twitter at any given point in time. This API enables easy access to relevant keywords and hashtags that constitute the bulk of Twitter chatter at any given moment.

4.1.3 On-demand APIs: Constraints and Proposed Workarounds

Internal API Inconsistencies: Known issues

Historical differences in the design and programming of the two APIs (as discussed previously in Section 4.1.2) led to inconsistencies between similar metadata fields in different parts of Twitter's internal API implementations [Russell, 2011a; Twitter Inc., 2011a, 2012a].

The main problem was due to a "long-lived bug with the Twitter API's `/search` resource" [Russell, 2011a]; which is a known issue since 2008[2]. A user `id` extracted from

[1] Although the 'REST `search` API' is, in theory, a Representational State Transfer API, naming convention ignores the REST prefix [Twitter Inc., 2012a].

[2] A forum discussion on Google Code reported the disparity of user IDs from different APIs, as early as December 2008: <`http://code.google.com/p/twitter-api/issues/detail?id=214`>.

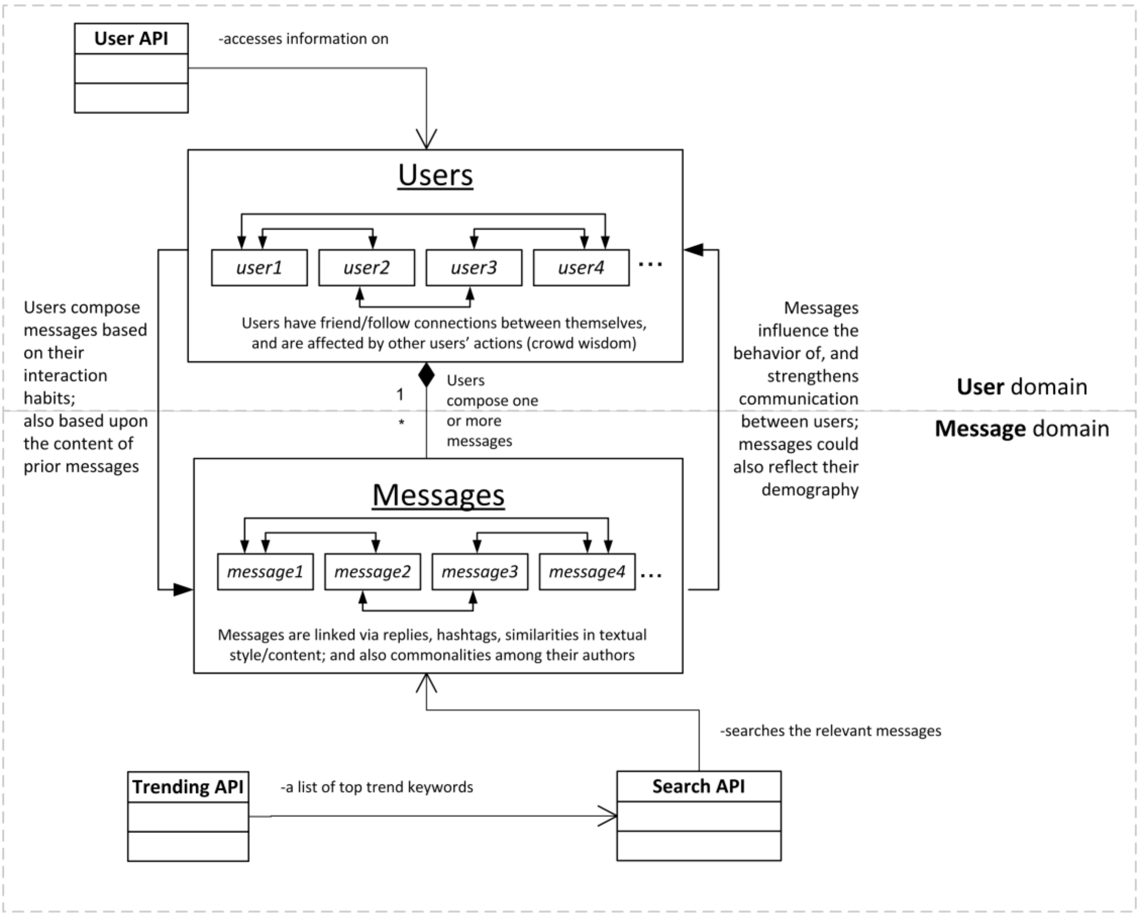

User API
-accesses information on
Users
user1
user2
user3
user4
...
Users have friend/follow connections between themselves, and are affected by other users' actions (crowd wisdom)
Users compose messages based on their interaction habits; also based upon the content of prior messages
1
*
Users compose one or more messages
Messages influence the behavior of, and strengthens communication between users; messages could also reflect their demography
User domain
Message domain
Messages
message1
message2
message3
message4
...
Messages are linked via replies, hashtags, similarities in textual style/content; and also commonalities among their authors
-searches the relevant messages
Trending API
-a list of top trend keywords
Search API

Figure 4.1: An annotated UML interaction diagram illustrating the interdependence of the two APIs, and consequently the two domains of users and messages, circa 2009.

messages obtained from the `search` API will not correspond to user `id` values "...in other APIs, such as the various `user` resources" [Russell, 2011a].

Internal API Inconsistencies: Workarounds

Hence, a workaround would be to use some other form of uniquely identifying metadata in the results from `search`, such as `from_user` as recommended bv Twitter Inc. [2012a] and used in [Cheong and Lee, 2009, 2010c; Russell, 2011a]. This workaround, albeit simple, is not a reliable unique identifier for a user, as a user can frequently change her username on Twitter. A user ID on the other hand, is more unique and constant.

This issue has finally been resolved; "...as of Nov 7, 2011 the Search API returns Twitter user IDs that match the Twitter REST API" [Twitter Inc., 2012a]. However, the problem remains for old or legacy data sets harvested using said API before November 2011; examples of such legacy data sets include the ones used in my early research [Cheong and Lee, 2009, 2010c].

Rate-limiting: Known issues

Due to technical limitations, the Twitter `search` API returns only a limited amount of tweets matching a user query. Two conditions affect this limitation, viz.:

1. A hard upper bound of 1500 tweets for a given batch of search results. I have identified this limit as early as 2009 [Cheong and Lee, 2009]; subsequently, Russell [2011a] has independently checked that the limit was enforced as of January 2011. This limit is still present as of July 2012.

2. If the search result quota does not constrain the set of returned results, a soft limit then applies to the date range. This is approximately 20 days before the current day, as I have found from my empirical studies [Cheong and Lee, 2009, 2010c].

For user information harvested from the REST-`user` API, I was able to retrieve user metadata for up to a maximum of 20,000 users per hour during my initial research in 2009. This is allowed only after explicitly being granted *white-listing* permissions from Twitter Inc for research purposes [Cheong and Lee, 2010c].

However, since the 2010 World Cup event, which took place circa June-July 2010, the rate-limiting has been dynamically-adjusted[3]. by Twitter Inc, which further "[lowered] the read-load on the API" and removed API white-listing privileges for REST-`user` and `search` [Twitter Inc., 2012a].

As of this thesis's time of writing, "the reduced rate limit is proportional to that application or users' allowance rather than a fixed value of requests" [Twitter Inc., 2012a]. The current estimate is 350 requests per hour, with white-listing still disabled, which is unlikely to change in the near future. In short, at time of writing, a significantly reduced number of accesses per hour to the REST-`user` API is attainable at time of writing compared when research on this PhD first started in 2009. The `search` API quota still stands at 1500 accesses per hour [Russell, 2011a].

Rate-limiting: Workarounds

Having said that, however, several workarounds for the rate-limiting problem have been developed in related work. This is mainly present in papers authored pre-2011, when the usage of the old APIs was still commonplace (see Chapter 3, e.g. [Krishnamurthy et al., 2008; Kwak et al., 2010]), and when the new rate-limiting scheme was yet to be in existence:

- **Constrained Crawl with Sampling:** Earlier papers written in and before 2009 [Huberman et al., 2008a; Java et al., 2009; Krishnamurthy et al., 2008] restricted their scope of research to a limited, but representative sample that is significant enough for research purposes. In these papers, the authors performed a constrained crawl of users originating from a seed user, as their research focused mainly on user relationships. This is different from harvesting a set of users based on the message similarity (e.g. discussing a particular topic or hashtag in the message domain). Such a method of sampling was also proposed by Cormode et al. [2010] as a method of "random node identification" in social media analysis.

[3] As reported in the Twitter Developers' Rate Limiting FAQ, updated on 9 August 2011: <`https://dev.twitter.com/docs/rate-limiting/faq`>.

- **Distributed Processing:** Kwak et al. [2010] used parallel computing in their research to crawl the entire Twitter user base. The authors used a cluster of about 20 machines (with individual IP addresses), all of which were white-listed by Twitter Inc. to access the Twitter API for research purposes. By limiting each machine to 20k total API requests per hour to avoid going against the term of service, Kwak et al. [2010] managed to achieve a theoretical maximum of 400k User API requests per hour. They were also capable of harvesting ten Trending Topics and 1500 tweets every five minutes. This method of distributing the load of querying the Twitter API in parallel with multiple clients is an efficient way to overcome Twitter Terms of Service restrictions and obtain a near-complete set of data. The obvious disadvantage of this approach was the high cost and the large amount of resources needed.

- **Polling with User Sampling:** One proposed method of obtaining a near-complete stream of tweets matching a particular criterion is continuous polling of the `search` API, as proposed in a prototype of mine (Section 5.1, published as Cheong and Lee [2010c]) and used in my case study Cheong and Lee [2009]. As the maximum search results for tweets of a given topic are limited to 1500 items, the API is repeatedly "polled". Polling is done by periodically repeating the search query after a particular interval to achieve a near-continuous stream of data. As the maximum number of users that can be queried from the REST-`user` API is fixed at 20000 users per hour (as in 2009, after white-listing), the only workaround to this is by performing random sampling of users. The degree of sampling is highly variable based on the needs of individual experiments. Weaknesses of this method include missing message data if messages are produced quicker than they are being consumed; and the inconsistency between the total number of messages obtained versus the number of users due to rate limit differences.

- **Polling with Caching:** The *polling with user sampling* method has been improved using a set number of search operations with fixed intervals — approximately 10 minutes between each run — which lets me achieve a capacity of tens of thousands of messages per hour [Cheong, 2010]. The total of search operations invoked (search queries per interval multiplied by number of intervals per hour) corresponds with the maximum user data retrieval limit of 20k per hour as stated. To improve the speed of user data collection, a simple caching mechanism is used by saving user metadata in memory. If a particular user is seen in a future tweet, the user's data can simply be accessed from memory as opposed to invoking another API call which is redundant. The principle behind this is that users are likely to contribute more than one tweet during the observation period; based on studies conducted on the user base [Cheong and Lee, 2009], studies on communication patterns [Boyd et al., 2010; Honeycutt and Herring, 2009], and Twitter data science textbooks [Russell, 2011a,b; Makice, 2009b]. One positive side-effect from the cache mechanism is that anomalous user records will suddenly cease to be unavailable for access on the Twitter API; contrary to their presence in previously-retrieved messages. This is a result of Twitter Inc. banning such users due to terms-of-service violations, such as spreading malware or spam, cf. [Cheong and Lee, 2009; Thomas et al., 2011]. The cache mechanism keeps track of the count of such confirmed spam accounts. (Related spam-detection heuristics were discussed prior in Section 3.3.4).

4.1.4 REST API: Other Resources and Features

Besides the REST-`user` resource, several other resources — sub-APIs under the REST API banner — can be used by developers and researchers to obtain metadata on other features of Twitter (most of which were introduced post-2009). For the sake of completeness, I briefly describe these other resources, parts of which will be revisited in Section 4.3.

- **Lists API**
  - Announced in September 2009, implemented in October 2009 [Twitter Inc., 2011b].
  - Lists are "collections of tweets, culled from a curated list of Twitter users" [Twitter Inc., 2012a]. The purpose of lists is to allow any Twitter user to categorize/group particular users in a "compilation that makes sense." [Twitter Inc., 2011b]
  - According to Twitter Inc, usage suggestions include "a list of the funniest Twitter accounts of all time, athletes, local businesses, [or] friends" [Twitter Inc., 2011b].
  - A list can be created either for public subscription, or set as private to its creator as a way to group connected users (see also Table 3.2 in Section 3.1.1 with relevant discussion from [Krishnamurthy, 2009]).
- **Retweeting API**
  - Announced in August 2009, implemented in November 2009 [Twitter Inc., 2011b].
  - This addition to the Twitter API is to allow "efficient dissemination of information across the entire Twitter ecosystem" [Twitter Inc., 2011b], and a way to "formalize retweeting by officially adding it to... [the Twitter API] platform and Twitter.com." [Twitter Inc., 2011b].
  - This addition to the Twitter platform resulted in an improved retweeting interface on the Twitter website (e.g. displays of retweets in a user's timeline, additional options to easily retweet a particular message), and the addition of several metadata items in the message domain specifically dealing with retweets (Section 4.3.1).
- **Places & Geo API**
  - Announced and implemented in August 2009 [Twitter Inc., 2011b].

  - As it stands, there are weaknesses with the user profile `location` text field (see Section 4.3.2). As "anything can be written in this field, [making it] not very dependable," [Twitter Inc., 2011b] hence, developers at Twitter have planned an improved API to "allow developers to add latitude and longitude to any tweet" [Twitter Inc., 2011b].
  - Hence, the Places & Geo API have been created to support per-message location information[4], as well as for the discovery of location-specific tweets & and associated data on locations. Such geographic metadata will be discussed in Section 4.3.2).

- **Suggested Users API**
  - Announced and implemented January 2010 [Twitter Inc., 2011b].
  - Developers at Twitter Inc have "created a number of algorithms to identify users across a variety of clusters who tweet actively and are engaged with their audiences" [Twitter Inc., 2011b].
  - Via the Suggested Users API, a Twitter user can obtain suggestions of popular users to follow.
  - The list of such users is algorithmically-selected by Twitter based on the areas of interest of the requesting user.

- **Trends API** and **Local Trends API**
  - The original Trends API — historically part of the `search` API; but since moved to the REST API — is a simple API that reveals the top ten Trending Topics discussed on Twitter on a period of time based on Twitter's proprietary algorithms [Twitter Inc., 2012a; Cheong, 2009].
  - This API enables easy access to relevant keywords and hashtags which constitute the bulk of Twitter chatter at any given moment.
  - An enhanced version of Trends, the *Local Trends* API, announced and implemented in January 2010 [Twitter Inc., 2011b], allows the retrieval of trends "people are talking about... on the state and city level" [Twitter Inc., 2011b] as opposed to the entire Twitter user base in general.
  - However, despite their similarities, Local Trends uses the *Yahoo! Where On Earth ID* to specify geographic regions; it has a different *modus operandi* from the Places & Geo API above, as of time of writing.

[4] According to Raffi Krikorian from the Twitter Development team, "for [the] first pass, we're only going live with United States-centric data, but that will quickly be expanded geographically as we work out the kinks in our system": <`http://groups.google.com/group/twitter-api-announce/browse\_thread/thread/e7fc06e4a8cb7150`>.

4.2 Twitter's Streaming API

4.2.1 Overview of the Streaming API

The new Streaming API, which complements the existing REST-`user` and `search` APIs, was launched in late 2009. Initially, little attention has been paid to the Streaming API; a lot of published research in 2009-2010 — [Cheong and Lee, 2009; Huberman et al., 2008a; Java et al., 2009; Kwak et al., 2010; Starbird et al., 2010] to name a few — still depended on message and user metadata using the on-demand `search` and REST-`user` APIs.

As the development of the Streaming API matured, I began to assess the viability of using the Streaming API as opposed to the on-demand APIs for my research.

In late 2010, I decided to use the Streaming API as a viable alternative for data collection, for the following reasons, mostly discussed in detail in Section 4.1.3:

1. **Existing weaknesses of the on-demand APIs**, specifically the inconsistencies between similar metadata fields in both APIs, and the need to query the REST `user` API to return user-domain metadata, as the `search` API only returns message metadata (as described in Section 4.1.3).

2. **Quotas and rate-limiting of the on-demand APIs** affected the amount of data I could retrieve for experimental purposes. The original limitations imposed in 2009 when research for this PhD thesis started (a maximum of 1500 messages per search query, up to 20k `user`s per hour after white-listing) were inconvenient, but did not pose a major hindrance in my research. However, Twitter Inc began to impose a dynamic but severely-limited quota in mid-2010, while discontinuing white-listing; allowing only hundreds of `user` queries per hour (as of time of writing). This further constraint made large-scale data collection practically infeasible.

The Streaming API, on the other hand, is capable of generating a very high number of metadata samples, without the imposition of rate-limits. Twitter Inc has favored the use of this API for data-collection purposes, as it can "provide useful low-latency samples without overwhelming clients or incurring excessive delivery cost" [Twitter Inc., 2012a].

Twitter Inc and its data reseller Gnip Inc, has recommended the use of the Streaming API for my research, according to F. Funke [pers. comm., 26 March 2011]. It is estimated to provide up to a maximum of ~1.4 million tweets per day, based upon the estimated daily volume of 140 million tweets per day. At time of writing, this API returns approximately "~1% of public statuses by default" [Twitter Inc., 2012a].

Furthermore, the Streaming API is more convenient than the on-demand APIs as it automatically embeds user metadata within the metadata of each message that it produces. This eliminates the need for a separate API call to access a users information, unlike the `search` API which only returned a `name` and an `id` associated with a message and requires a separate call to the REST-`user` API in order to fetch the user metadata (cf. my framework in Section 5.1 [Cheong and Lee, 2010c]).

4.2.2 Streaming API Concepts

Similar to the on-demand APIs, the Streaming API returns tweets designated as public, as opposed to private or hidden tweets. The API further filters messages "for quality", eliminating questionable tweets from "suspended accounts, or accounts that may jeopardize search quality" [Twitter Inc., 2012a]. Different from the on-demand (REST) APIs however, the Streaming API works using sockets: the user establishes a socket connection to the Streaming API prior to calling its methods. If the connection is successful, the API then continuously streams a sample of public tweets, along with its associated message and user metadata, encapsulated in JSON format. This process continues until the socket connection is terminated by the user or due to an error (such as overloading of the Streaming API, or a network error).

Due to the high volume of data that can potentially be accessible, Twitter Inc recommends the use of "decoupled collection, processing and persistence components" [Twitter Inc., 2012a] when designing a program that consumes data from the Streaming API:

> *...for example, collect "raw" statuses (that is, not parsed or marshaled into your language's native object format) in one process, and pass each status into a queueing system, rotated flatfile, or database. In a second process, consume statuses from your queue or store of choice, parse them, extract the fields relevant to your application, etc.* [Twitter Inc., 2012a]

To obtain a sample of tweets (from the set of all available tweets created at any given moment), the `sample` method is called from the Streaming API. As mentioned earlier, the Twitter Streaming API uses a sampling algorithm to produce approximately "~1% of public statuses by default" [Twitter Inc., 2012a].

This sampling algorithm, as of time of writing, works as follows [Twitter Inc., 2012a]:

> *...the status id modulo 100 is taken on each public status, that is, from the Firehose [Twitter's codename for the entire stream of tweets]. Modulus value `0` is delivered to Spritzer, and values `0-10` are delivered to Gardenhose [Twitter's codename for a paid service that returns ~10% of all public statuses]. Over a significant period, a 1% and a 10% sample of public statuses is approached. This algorithm, in conjunction with the status id assignment algorithm, will tend to produce a random selection.*

The Streaming API's `filter` method can also be invoked to narrow the stream to a smaller subset of tweets that match a certain criteria [Twitter Inc., 2012a]. Of interest is the `track` parameter, allowing the retrieval of tweets matching a particular search query string. I have authored a technical description of the Streaming API's inner workings, provided for reference in Appendix C.

4.2.3 Streaming API versus On-demand APIs

To sum up the discussion on the two kinds of APIs available on Twitter that are of use for research — on-demand versus streaming — I provide here a summary of the differences between the two APIs in Table 4.1.

Feature	**On-demand APIs:** `search` **and REST-**`user`	**Streaming API**
Connection	API methods are called as required.	Socket connection needs to be established.
Result format	Data is returned in a representational manner; connection is ended once data is returned.	Data will be streamed through the open socket, until socket is explicitly closed.
Range of results	Returns results up to the time of API call.	Returns results only after the socket is opened successfully.
Number of results	A maximum of 1,500 tweets per query; a dynamic amount (~350) users per hour (Section 4.1.3).	A maximum of ~1% of all tweets; each with its linked user metadata (~1.4 million daily).
Research implications	Used widely in research prior to 2011; main obstacle is the current rate-limit.	Viable alternative for large-scale data collection

Table 4.1: Summary comparison of the on-demand APIs (`search` and REST `users`) and Streaming API.

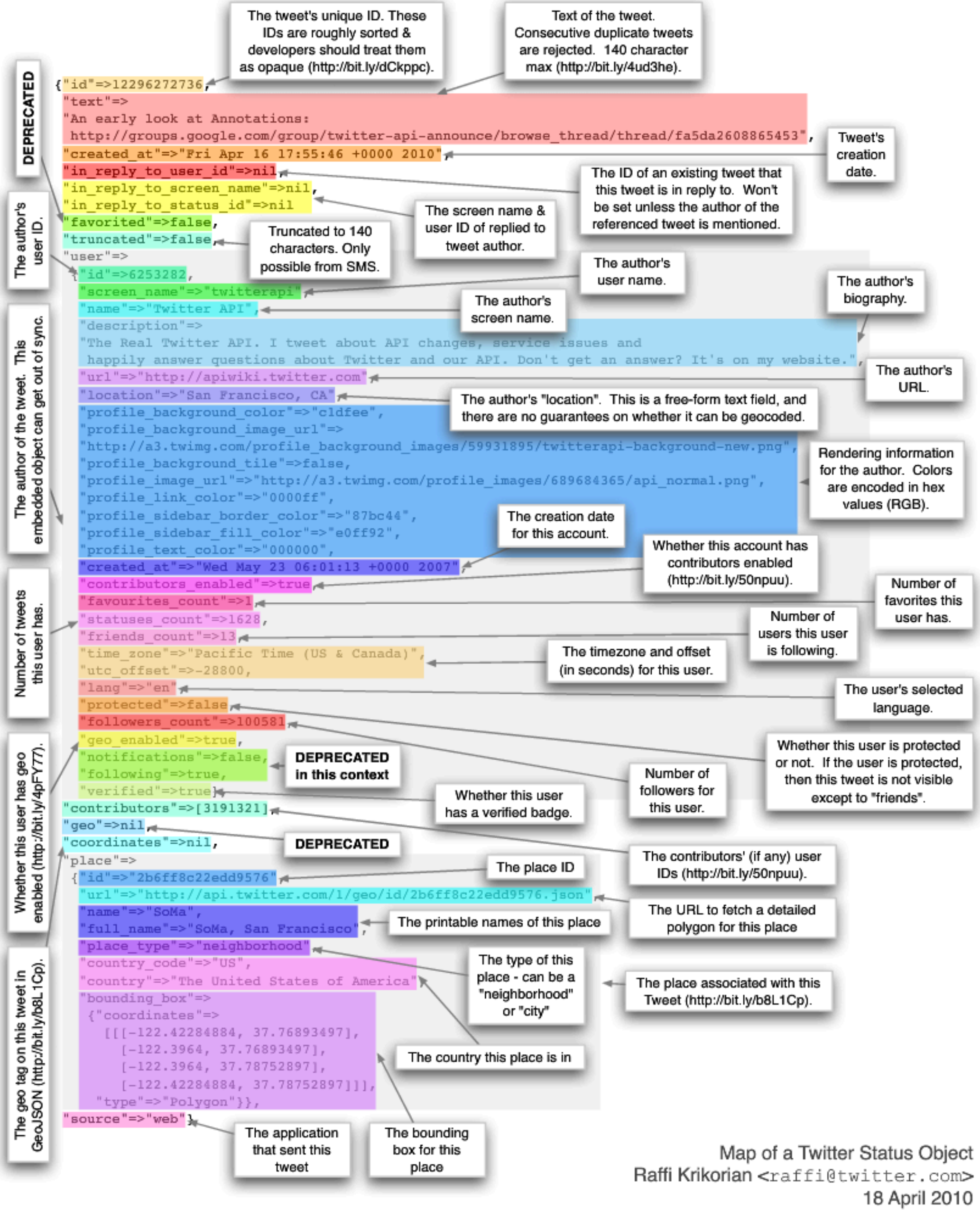
4.3 Metadata in Twitter Domains

In this section, I discuss the raw metadata that can be obtained from the user and message domains via the Twitter APIs [Makice, 2009b], as of time of writing. Figure 4.2 is a graphical overview of the complete set of raw Twitter metadata available, as illustrated Krikorian [2010], working for Twitter Inc.

For the sake of completeness, an in-depth technical explanation of every available metadata field returned from the Twitter API (as of time of writing), as well as sample raw metadata, are provided in Appendix A. , For brevity, in this section I only enumerate metadata items that are featured in my research contributions. The metadata in the next two subsections are discussed as-is. Their potential applications or uses will be covered by my research in Section 4.6 onwards.

4.3.1 Message Domain

Table 4.2 itemizes four useful metadata items that can be found within the message domains.



```
{"id"=>12296272736,
"text"=>
"An early look at Annotations:
 http://groups.google.com/group/twitter-api-announce/browse_thread/thread/fa5da2608865453",
"created_at"=>"Fri Apr 16 17:55:46 +0000 2010",
"in_reply_to_user_id"=>nil,
"in_reply_to_screen_name"=>nil,
"in_reply_to_status_id"=>nil
"favorited"=>false,
"truncated"=>false,
"user"=>
 {"id"=>6253282,
  "screen_name"=>"twitterapi",
  "name"=>"Twitter API",
  "description"=>
  "The Real Twitter API. I tweet about API changes, service issues and
   happily answer questions about Twitter and our API. Don't get an answer? It's on my website.",
  "url"=>"http://apiwiki.twitter.com",
  "location"=>"San Francisco, CA",
  "profile_background_color"=>"c1dfee",
  "profile_background_image_url"=>
  "http://a3.twimg.com/profile_background_images/59931895/twitterapi-background-new.png",
  "profile_background_tile"=>false,
  "profile_image_url"=>"http://a3.twimg.com/profile_images/689684365/api_normal.png",
  "profile_link_color"=>"0000ff",
  "profile_sidebar_border_color"=>"87bc44",
  "profile_sidebar_fill_color"=>"e0ff92",
  "profile_text_color"=>"000000",
  "created_at"=>"Wed May 23 06:01:13 +0000 2007",
  "contributors_enabled"=>true
  "favourites_count"=>1
  "statuses_count"=>1628,
  "friends_count"=>13
  "time_zone"=>"Pacific Time (US & Canada)",
  "utc_offset"=>-28800,
  "lang"=>"en"
  "protected"=>false
  "followers_count"=>100581
  "geo_enabled"=>true,
  "notifications"=>false,
  "following"=>true,
  "verified"=>true},
"contributors"=>[3191321],
"geo"=>nil,
"coordinates"=>nil,
"place"=>
 {"id"=>"2b6ff8c22edd9576",
  "url"=>"http://api.twitter.com/1/geo/id/2b6ff8c22edd9576.json",
  "name"=>"SoMa",
  "full_name"=>"SoMa, San Francisco",
  "place_type"=>"neighborhood",
  "country_code"=>"US",
  "country"=>"The United States of America",
  "bounding_box"=>
   {"coordinates"=>
     [[[-122.42284884, 37.76893497],
       [-122.3964, 37.76893497],
       [-122.3964, 37.78752897],
       [-122.42284884, 37.78752897]]],
   "type"=>"Polygon"}},
"source"=>"web"}
```

The tweet's unique ID. These IDs are roughly sorted & developers should treat them as opaque (http://bit.ly/dCkppc).

Text of the tweet. Consecutive duplicate tweets are rejected. 140 character max (http://bit.ly/4ud3he).

Tweet's creation date.

DEPRECATED

The ID of an existing tweet that this tweet is in reply to. Won't be set unless the author of the referenced tweet is mentioned.

The screen name & user ID of replied to tweet author.

The author's user ID.

Truncated to 140 characters. Only possible from SMS.

The author's user name.

The author's screen name.

The author's biography.

The author of the tweet. This embedded object can get out of sync.

The author's URL.

The author's "location". This is a free-form text field, and there are no guarantees on whether it can be geocoded.

Rendering information for the author. Colors are encoded in hex values (RGB).

The creation date for this account.

Whether this account has contributors enabled (http://bit.ly/50npuu).

Number of favorites this user has.

Number of tweets this user has.

Number of users this user is following.

The timezone and offset (in seconds) for this user.

The user's selected language.

Whether this user is protected or not. If the user is protected, then this tweet is not visible except to "friends".

Number of followers for this user.

Whether this user has geo enabled (http://bit.ly/4pFY77).

DEPRECATED in this context

Whether this user has a verified badge.

The contributors' (if any) user IDs (http://bit.ly/50npuu).

DEPRECATED

The geo tag on this tweet in GeoJSON (http://bit.ly/b8L1Cp).

The place ID

The URL to fetch a detailed polygon for this place

The printable names of this place

The type of this place - can be a "neighborhood" or "city"

The place associated with this Tweet (http://bit.ly/b8L1Cp).

The country this place is in

The application that sent this tweet

The bounding box for this place

Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

Figure 4.2: A Twitter API developer's overview of raw Twitter metadata [Krikorian, 2010], at <http://datasift.com/a/wp-content/themes/datasift/images/tweet_diagram.pdf>.

Table 4.2: Metadata items in the message domain, together with a brief explanation of its role in Twitter.

Name	**Explanation**
`text`	The raw message text (up to 140 characters), the most visible attribute of a message. Substrings of this could take the form of URLs, hashtags, `@user` references, and 'smileys' such as "`:)`"
`id`	The unique message identifier for each tweet.
`source`	Identifier for the software used to publish a particular tweet, which can be either be official Twitter services (`web, mobile, txt` for the main Twitter website, mobile site, and SMS interface respectively), or third-party applications with a hyperlink to their official website.
`created_at`	Time-stamp indicating when the tweet was composed.
`retweet_count`	Introduced as part of the Retweeting API (Section 4.1.4), this value stores the number of Retweets (RTs) that the current message has. However, if the amount of retweets for a message exceeds the order of a hundred, `retweet_count` will instead be the constant string `100+`.

4.3.2 User Domain

As for the user domain, Table 4.3 itemizes thirteen useful metadata items that can be found within the message domains. These items were used in my research [Cheong and Lee, 2010a; Cheong and Ray, 2011]; and also the work of others as described in my earlier literature review (Chapter 3).

4.4 Preliminary Investigation on Metadata Interdependence

Research performed in the course of this PhD has identified several real-world properties that can be inferred from metadata in both user and message domains. A preliminary investigation on the connections between metadata in the two Twitter domains in the early days of my PhD research (late 2009–early 2010) briefly summarized the possible features that can potentially be inferred.

From my literature reviews [Cheong and Lee, 2010a; Cheong and Ray, 2011] in Chapter 3, I have identified several areas lacking in existing research, which is illustrated in Figure 4.3.

The following list elaborates further on the annotations covered in Figure 4.3, which in turn was based on the earlier UML interaction diagram in Figure 4.2 illustrating the interdependence of the two APIs.

- Existing studies on properties and emergent features [Java et al., 2009; Krishnamurthy et al., 2008; Huberman et al., 2008a] cover a bit of both domains, and serve as a seed point for my research.

- Studies on message addressivity and forwarding [Honeycutt and Herring, 2009; Boyd et al., 2010], and messaging in times of crisis and convergence [Starbird et al., 2010;

Table 4.3: Metadata items in the user domain, together with a brief explanation of its role in Twitter.

Name	**Explanation**
`id`	The unique identifier for each user, similar to its namesake in the message domain.
`name`	This contains the real name of a user.
`screen_name`	The username or Twitter account name for a Twitter user, this is frequently denoted with a `@` suffix in tweets.
`location`	User-provided, 30-character limited text string for a user to describe his/her current location. This can be free-form text naming a location, or the exact geographical coordinates as generated by GPS-enabled Twitter clients.
`URL`	User website as published in their profile.
`statuses_count`	Total number of statuses created by a user, since his/her Twitter account was created.
`created_at`	Time-stamp generated during the creation of the user's Twitter account.
`listed_count`	Number of lists the Twitter user belongs to, as added by other users. This variable was introduced as part of the deployment of the Lists API (Section 4.1.4).
`followers_count`	The number of other Twitter users currently *following* the user (node in-degree).
`friends_count`	The number of other users the current user *follows* (node out-degree).
`default_profile`	Boolean value indicating if the Twitter user profile is the default style (uncustomized) or otherwise.
`default_profile _image`	Boolean value indicating if the user has customized his/her profile picture, or left it to the default version.
`verified`	Boolean value identifying if a user is "*Twitter Verified*", i.e. a high-profile user, such as a celebrity or politician, who has applied to Twitter Inc for identity verification.

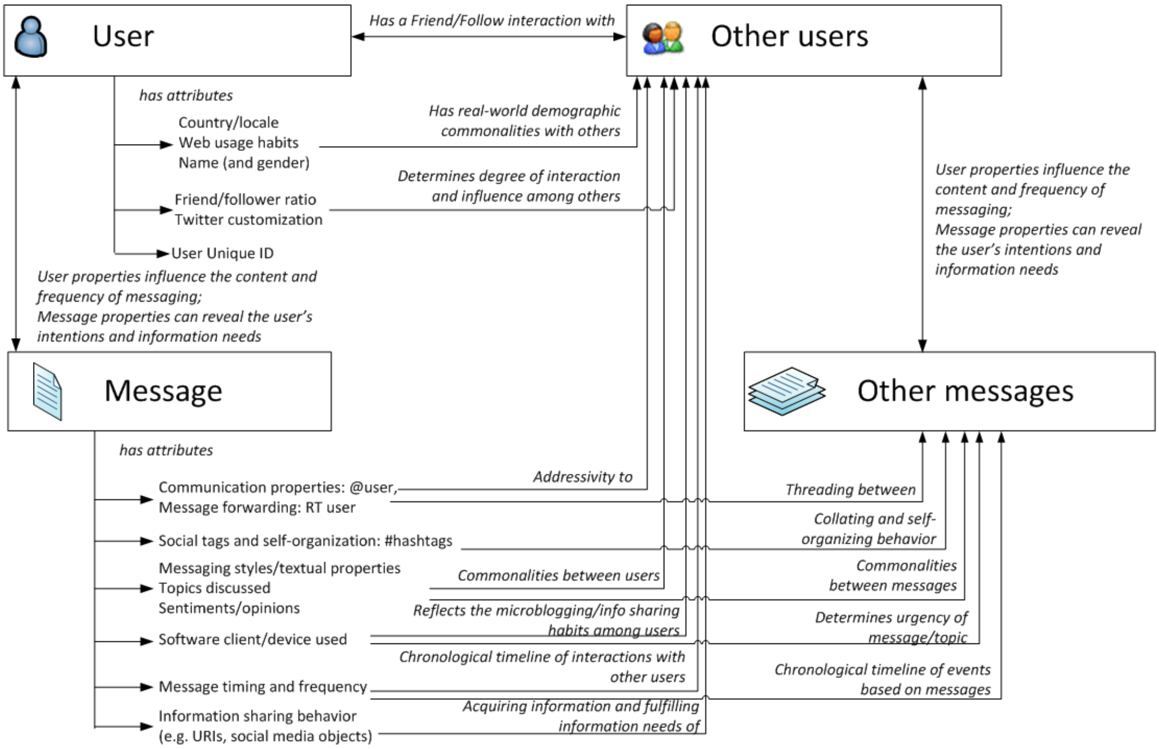

User
Has a Friend/Follow interaction with
Other users
has attributes
Country/locale
Web usage habits
Name (and gender)
Has real-world demographic commonalities with others
Friend/follower ratio
Twitter customization
Determines degree of interaction and influence among others
User Unique ID
User properties influence the content and frequency of messaging;
Message properties can reveal the user's intentions and information needs
User properties influence the content and frequency of messaging;
Message properties can reveal the user's intentions and information needs
Message
Other messages
has attributes
Communication properties: @user,
Message forwarding: RT user
Addressivity to
Threading between
Social tags and self-organization: #hashtags
Collating and self-organizing behavior
Messaging styles/textual properties
Topics discussed
Sentiments/opinions
Commonalities between users
Commonalities between messages
Reflects the microblogging/info sharing habits among users
Software client/device used
Determines urgency of message/topic
Chronological timeline of interactions with other users
Message timing and frequency
Chronological timeline of events based on messages
Acquiring information and fulfilling information needs of
Information sharing behavior
(e.g. URIs, social media objects)

Figure 4.3: Connections between the user and message domains from early 2010, which gave rise to the study of emergent features and potential inferences from raw metadata in this chapter.

Hughes and Palen, 2009; Sutton et al., 2008] gave emphasis to the message domain, with only a brief analysis of the user domain.

- Approaches to trend analysis from blogs and other social media [Gruhl et al., 2004; Fukuhara et al., 2005; Gruhl et al., 2005] considers the message domain, specifically the chronological distribution of tweets to predict spikes and parallels with real-world activity. These concepts were ported to Twitter in future work such as Kwak et al. [2010].

- Sentiment and opinion analysis on Twitter [Jansen et al., 2009a; Banerjee et al., 2009; Shamma et al., 2009] ties in properties from both the user and message domains. Sentiment and opinion analysis obviously take place in the message domain; augmented by a limited amount of user information (such as locale/countries and messaging habits).

Initially, before the development of the Twitter Streaming API (Section 4.2.1) has matured, I experimented with small-scale data sets made available via the Twitter REST (on-demand) APIs. The number of records range from an order of tens, up to a magnitude of thousands of results [Cheong and Lee, 2009, 2010c,d]. This is compounded further by the need to look up user metadata separately from the message metadata (discussed in Section 4.1.3), causing inconsistencies in the number of complete records. Also, another difficulty in experimenting on my original data set was the lack of consistency in sampling, as the REST APIs contain rate limits, making it suitable only for looking up specific tweets, but ill-suited for capturing data consistently over time.

4.5 Introducing the *10-Gigabyte Dataset*

For the purposes of testing and evaluation of the algorithms in this thesis, there is a need for a complete and representative dataset of real-world tweets, which naturally includes complete user and message metadata records.

I introduce the *10-Gigabyte Dataset*, consisting of 7,863,650 tweets (with complete message metadata), from 4,491,022 unique users (again with complete user metadata). This dataset was gathered in November 2011 from real-world tweets, sourced from the Streaming API after some discussion on its suitability with Twitter Inc's authorized data reseller, Gnip Inc. [F. Funke, pers. comm., 26 March 2011].

An in-depth discussion on the *10-Gigabyte Dataset*, including its properties, data collection techniques, idiosyncrasies, as well as the prototype used in collecting the data is located in Section 5.3. In terms of the current chapter's coverage, the brief aforementioned summary should suffice.

4.6 Learning by Inference: Real-life Demographic Properties

4.6.1 Gender

In prior literature [Jones et al., 2007], gender is identified as one of the attributes that has "subtle cues to [a user's] identity". Work by Joinson [2008] and Schrammel et al. [2008], for instance, have identified gender as one of the differentiating factors in influencing information sharing and social networking behaviour.

However, as discussed earlier in the definition of online social network characteristics (Table 3.2), Twitter has only a subset of profile features as compared to other *de facto* online social networks such as MySpace, Facebook, or Google+. Due to this limitation, Twitter has no facility to allow users to enter their gender into their profile information.

Previous research [Cheong and Lee, 2009] has identified the fact that users on Twitter frequently publish their names as opposed to an alias or nickname as part of their user information. To overcome the absence of a user's gender in Twitter profile information, I hypothesize that the gender of a Twitter user can be inferred by using a user's real name (the `name` metadata entry).

Research on Name-based Gender Detection

There are many known methods to identify a person's gender from her name; examples of which encompass research areas such name phonology [Slater and Feinman, 1985], and computer-based pattern recognition [Gallagher and Chen, 2008]. The paper by Gallagher and Chen [2008] is notable for using statistics on popular names released by the United

States Social Security Administration (US SSA) [U.S. Social Security Administration, 2011]. In the US SSA dataset, the most popular names used to register births in the United States for each given year are recorded since the year 1880.

Independently, Warden [2011] has published the *Data Science Toolkit* — a "specialized Linux distribution" [Warden, 2011] of tools related to data science. Within this toolkit, Warden [2011] has developed a *Text to People* API based on the Perl module `Text-GenderFromName-0.33` by Daly and Orwant [2003]. The API (and base module) allows researchers to "...[spot] text fragments that look like people's names or titles, and guesses their gender where possible" [Warden, 2011]. Again, the developers Warden [2011]; Daly and Orwant [2003] use raw data from the US SSA's "Most Popular Names of the 1980's" list of 1,001 male first names and 1,013 female first names" [U.S. Social Security Administration, 2011]. Their algorithm detects gender from a name string based on "...based on [name] exclusivity, [frequency-based] weight, metaphones... [and simple regular expression]-style matching" [Daly and Orwant, 2003].

Using Frequency Ranking of Real Names to Determine Gender on Twitter

Independent of the studies cited above, my initial study of gender detection in Twitter involves a simple ranking algorithm to determine the gender of a person based on statistics released by the United States Government [Cheong and Lee, 2009]. My initial prototype name ranking algorithm [Cheong and Lee, 2010c] used the United States Census department's `dist.female.first` and `dist.male.first` data sets [U.S. Census Bureau, 2010] as the ranking data. This dataset was created by the United States Census department based on 1990 raw census data, involving 6,188,353 total first names (after data sanitization) from a diverse range of ethnicities, sexes, and ages.

My algorithm differs from existing ones [Warden, 2011; Daly and Orwant, 2003] in that I only perform simple string matching as opposed to a hybrid (exclusivity, weight, and metaphone-based) algorithm. Instead, the simplicity and processing speed of a simple ranking algorithm is preferred, to account for a potentially high number of input name data.

The end result was a total of 5,494 unique and ethnically-diverse first names, separated into two ranked lists of male and female names, were used to statistically determine the gender based on a user's real given name. A hashing algorithm is first employed to preload all the first names on the census data for both males and females. When name string is encountered in Twitter's `name` metadata entry, it is first sanitized by removing all non-alphabetic characters. The first name is extracted, before its rank is looked up from the hash tables of male and female name frequency. The gender of the queried name is inferred based on this frequency information [Cheong and Lee, 2010c]. My approach is summarized in Algorithm 4.1.

To describe the inner workings of Algorithm 4.1, I qualitatively evaluated the outputs generated from several examples of input data:

- The name *Susan* is inferred to be *female*, as *Susan* has a rank of eight on the female name list, but is not found on the male name list.

Algorithm 4.1 Proposed gender algorithm based on frequency ranking.

```
procedure GenderFromName(firstname)
Require: malehash ← male-name frequency rankings from 1990 US Census data, indexed
    by name
Require: femalehash ← female-name frequency rankings from 1990 US Census data,
    indexed by name
    if name defined in malehash and name undefined in femalehash then
        return male
    else if name defined in femalehash and name undefined in malehash then
        return female
    else if malehash{rank} better than femalehash{rank} then
        return male
    else if femalehash{rank} better than malehash{rank} then
        return male
    else
        return indeterminate
    end if
end procedure
```

- The name *Dorian* is inferred to be *male*, as *Dorian* has a rank of 870 on the male name list; however it has a far lower rank of 2165 on the female name list. Similar to real life, Dorian is mainly a male name, but is also used (rather infrequently) as a female name.
- The name *Twitter* cannot be determined to be either male or female, as it is a proper name of a non-human entity instead.

To measure the accuracy of Algorithm 4.1, I conducted validation testing to compare the accuracy of human evaluation versus algorithmic gender determination (Experiment 4.1).

Experiment 4.1. *To validate the accuracy of gender inferences using Algorithm 4.1.*

Method: Algorithm 4.1 is run against ten sets of 100 names each, harvested at random from a set of 1,000 Twitter messages [Cheong and Lee, 2010c]. (For reference, this test set of 1,000 messages will be labeled as *FirstNameTestSet-2009*). For comparison, the ground truth is obtained by using a human volunteer to determine the genders.

Results and Discussion: The results are depicted in Table 4.4. Averaging the accuracy rates over each of the ten test sets, an average accuracy rate of approximately 86.6% [Cheong and Lee, 2010c] is attained.

The comparison obtained is based on the underlying assumption that human (manual) detection always represents the ground truth, i.e. detecting a person's gender based on name will be no problem for human testers.

It is pertinent to note that the names used in this test [Cheong and Lee, 2010c] were extracted in mid-2009, when the user base was estimated to be 75 million users[5]. This

[5] Estimated user base as of the end of 2009, by RJMetrics Inc.: <http://info.rjmetrics.com/blog/bid/44962/New-Data-on-Twitter-s-Users-and-Engagement>.

Test set	1	2	3	4	5	6	7	8	9	10
Performance: percentage of correctly determined genders	81%	88%	89%	82%	90%	83%	85%	92%	88%	88%

Table 4.4: Gender detection: algorithm accuracy (versus ground truth) using the US 1990 Census rankings, tested on *1k-FirstNames-2009*

figure increased to about 200 million as of 2011[6], with people from a diverse range of cultures and languages contributing to an almost 167% in user growth.

This is especially relevant, as the majority of the names tested against this algorithm consist of Western first names. Ideally first names from a diverse range of cultures and languages would need to be analyzed.

Improving Gender Detection using the US Social Security Baby Names Database

Inspired by Gallagher and Chen [2008]; Warden [2011]; Daly and Orwant [2003], and the ability to handle a wide variety of first names from different cultures, I experimented with the US SSA first names dataset [U.S. Social Security Administration, 2011] as an alternative to the 1990 US Census name data for my simple ranking algorithm.

In my preliminary analysis of the US SSA dataset, I found that it is more thorough as it covers first names from a wide variety of cultures (not merely limited to common Western first names), and more up-to-date than the 1990 Census data [U.S. Census Bureau, 2010]. Differing from Gallagher and Chen [2008]'s method for learning gender and age priors from the SSA dataset, however, I instead adapted my ranking system to use all 130 years worth of data on popular first names, from 1880–2010 (inclusive). To perform such adaptation, I summed the rank data for each name for a particular gender across 30 years.

In my improved algorithm, the raw data comes in the form of one comma-separated (CSV) file per year, where each record is formatted as: "`name, gender, frequency`". Total frequencies for each unique first name is then stored by using two hash tables - one for males, another for females - and indexed them by first name [Cheong et al., 2012b].

The resulting ranking data, adapted from of the SSA raw data [U.S. Social Security Administration, 2011], has totals of:

- 36,742 unique male names (from a total of 162,412,587 recorded male births)
- 61,406 unique female names (from a total of 159,990,140 recorded female births)

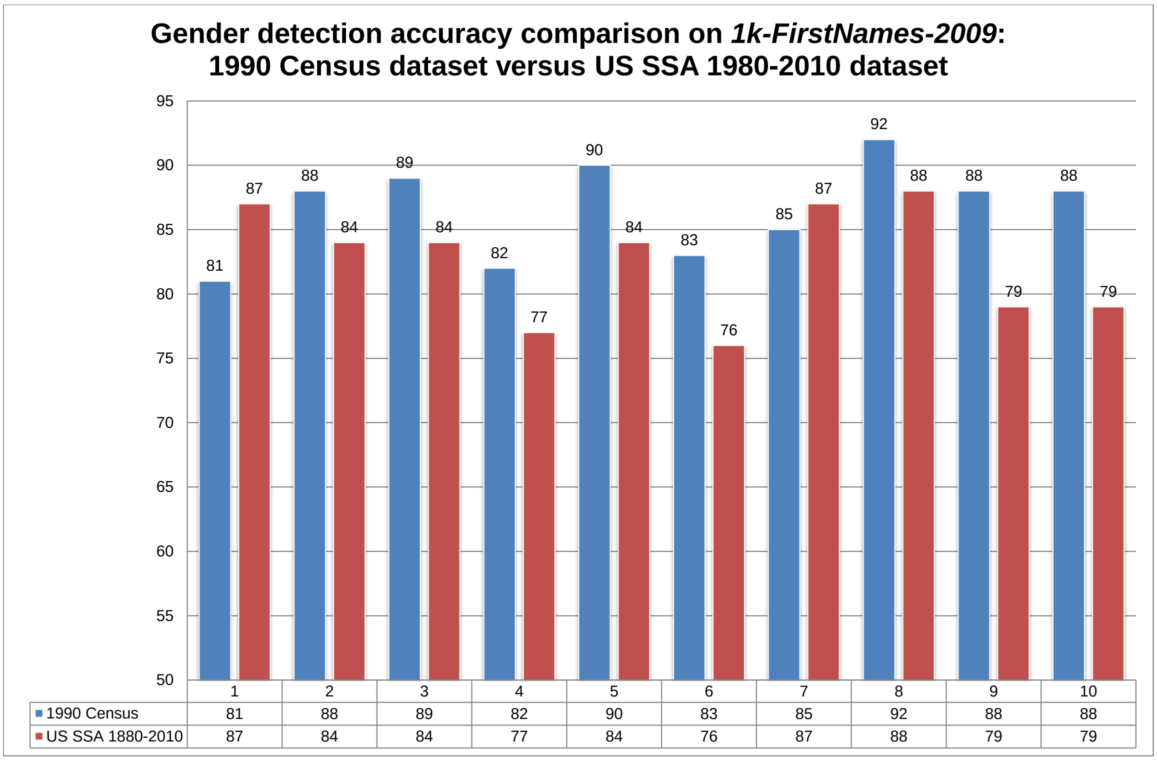
As the ranking dataset has changed in terms of quality and quantity of entries, the performance of the augmented algorithm has to be reevaluated against its predecessor (Experiment 4.2).

Experiment 4.2. *To validate the accuracy of gender inferences using the SSA dataset coupled with Algorithm 4.1.*

[6] Estimated user base as of the beginning of 2011, by Kathryn Corric.: <`http://kathryncorrick.co.uk/2011/02/17/the-state-of-the-twittersphere-in-february-2011/`>.

Method: Algorithm 4.1 (augmented with the new SSA dataset) is applied on *1k-FirstNames-2009* test data. The rest of the experiment is similar to Experiment 4.1.

Results and Discussion: The bar chart in Figure 4.4 compares the accuracy of Algorithm 4.1 (augmented with SSA ranking data) to the original algorithm (using 1990 Census ranking data). The average accuracy rate from using the SSA 1880-2010 ranking data is 82.5%, a slight drop compared to the 1990 US Census ranking data.



	1	2	3	4	5	6	7	8	9	10
■ 1990 Census	81	88	89	82	90	83	85	92	88	88
■ US SSA 1880-2010	87	84	84	77	84	76	87	88	79	79

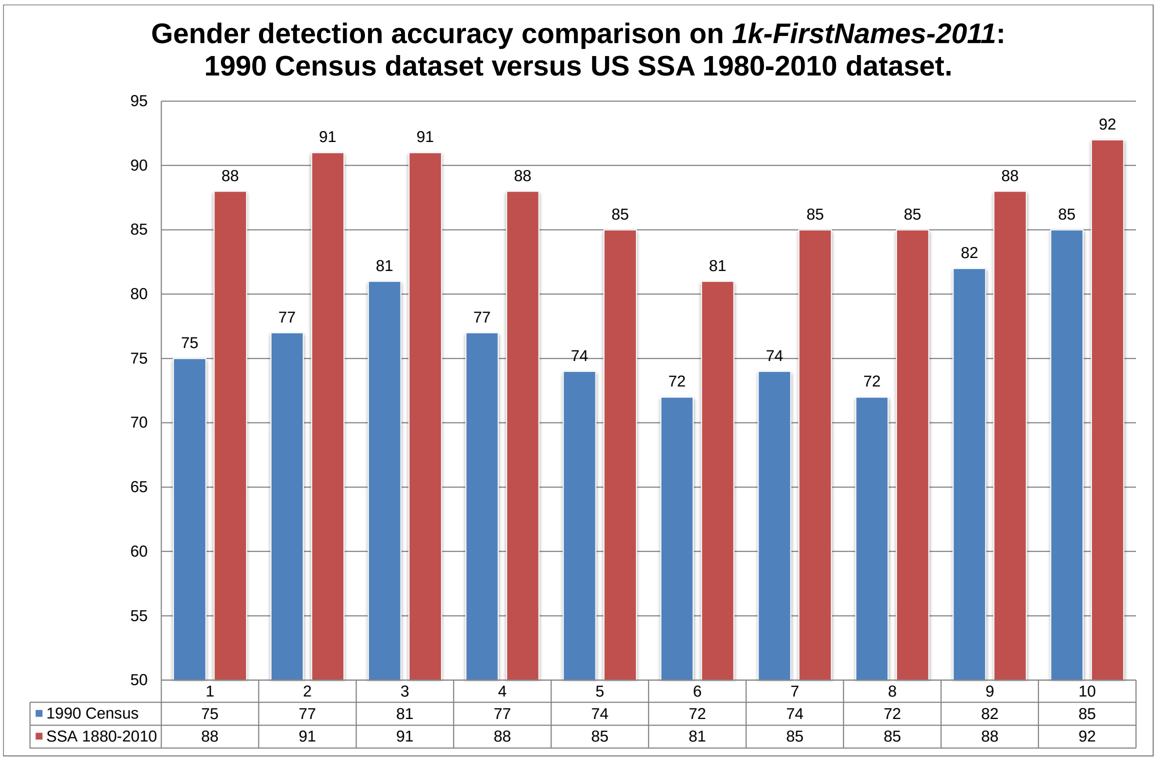
Figure 4.4: Comparison between ranking data used — 1990 Census versus SSA 1880-2010 — tested on *1k-FirstNames-2009*.

However, to further validate the proposed advantages of the SSA 1880-2010 ranking data and nullify the effect of Twitters evolving user base, I decided to rerun Experiment 4.2 on a more up-to-date test set of multicultural first names. This reevaluation is described in 4.3.

Experiment 4.3. *Comparing the accuracy of gender inferences obtained from Algorithm 4.1 (augmented with the new SSA ranking data), with the original algorithm (using 1990 US Census ranking data).*

Method: Another test set is created, using first names from current real-world Twitter users from a more diverse range of languages and cultures. Using the *10-Gigabyte Dataset* (Section 4.5 elaborates on the empirical data used), 1,000 multicultural first names are sampled into a new first name test set, *1k-FirstNames-2011*. Algorithm 4.1 with the 1990 Census ranking data is executed on [Cheong and Lee, 2010c] on *1k-FirstNames-2011*; this is subsequently repeated with SSA 1880-2010 ranking data.

Results and Discussion: Algorithm 4.1 augmented with the SSA dataset outperforms the earlier unaugmented version (which utilized the 1990 US Census rankings). Figure 4.5 illustrates the comparative results. The average accuracy obtained using the 1990 Census data is a mere 76.9%, compared to the augmented SSA 1880-2010 data which provides an average accuracy of 87.4%.



	1	2	3	4	5	6	7	8	9	10
🟦 1990 Census	75	77	81	77	74	72	74	72	82	85
🟥 SSA 1880-2010	88	91	91	88	85	81	85	85	88	92

Figure 4.5: Comparison between ranking data used — 1990 US Census (original Algorithm 4.1 versus SSA 1880-2010 (augmented Algorithm 4.1 — tested on *1k-FirstNames-2011*. The augmented version with SSA 1880-2010 ranking data clearly outperforms the original algorithm using 1990 US Census ranking data.

Evaluation

From Experiment 4.3, the usage of the SSA 1880-2010 ranking data to augment Algorithm 4.1 triumphed over the original algorithm with the 1990 US Census rankings, due to its up-to-dateness in reflecting the current *zeitgeist* in first name trends [Cheong et al., 2012b].

Based on said findings, my algorithm works successfully for common names. Several limitations that affect the accuracy of human validation (and by extension, algorithmic accuracy) have been identified. These include:

1. **Non-common names:** The algorithm is based on ranking data for common first names, and as such is not exhaustive; the same issue applies to humans as not all names will be familiar to a human tester, such as names from cultures the human tester is not familiar with. This issue is remedied to a certain extent by using the

latest first name ranking data [U.S. Social Security Administration, 2011]. Empirically, I have found that it works for non-Western names such as *Kareem* (Arabic), *Chan* (Chinese), and *Shigeru* (Japanese).

2. **Character sets:** this algorithm works with names that are presented using the Latin alphabet and is not capable of recognizing names from languages using another character set (such as the CJK character set for Chinese/Japanese/Korean characters), unless they are Romanized. In my experiments, I have only dealt with names spelt using the Latin alphabet and encoded in ASCII.

3. **Androgynous names:** names such as *Tracy*, *Kim* and *Lauren* are applicable for people of both genders; hence the algorithm (and even human testing) is not able to determine the gender accurately without other cues.

4. **Presence of names in non-human contexts:** if human names are present in non-human contexts, e.g. part of an organization's name, the algorithm does not ignore it as a human tester would. An example such a context are the names *Beverly Cinema* and *Orange UK*, where *Cinema* and *UK* are not surnames.

Compared to existing approaches, my proposed algorithm has the advantages of:

1. **Speed:** My algorithm only involves a simple hash-lookup operation which runs in constant-time (slightly sacrificing memory as a trade-off for time); this is beneficial for large-scale gender detection. When tested on an input of one million names, it outperfomed the `Text-GenderFromName-0.33` Perl algorithm [Daly and Orwant, 2003] — using parameters as recommended by the Perl documentation — by 8.64 seconds on average (107.7917 seconds versus 116.43838 seconds)[7].

2. **Adaptability:** If a new ranked set of gender data is available, such as future updates to the US SSA rankings, the algorithm can be trivially adapted to incorporate the updated data set.

As I have documented in my paper resulting from this study [Cheong and Lee, 2010c], as far as I know, this is the first time such name-based gender detection has taken place in the field of microblogging research.

4.6.2 Location

Location Hints and Cues in Twitter Metadata

In this thesis, I use geographic location as an important demographic property as it plays a role in studies such as the reach of online social networks based on geography (e.g. [Marsden, 2002]), and the dissemination of information during crisis and convergence (e.g.

[7] I arrived at these figures by obtaining the average of five runs per algorithm on an Intel Core Duo 3GHz CPU and 2GB of RAM running Windows XP. The repeated runs are to negate the influence of external factors such as CPU caching and background processes. I eliminated the issue of file fragmentation over the dataset by defragmenting the test data files, and ensure that Windows and background processes are not performing hard-disk intensive operations (such as paging) during the experiment.

[Starbird et al., 2010; Hughes and Palen, 2009; Cheong and Lee, 2010d]). As discussed in Section 4.3.2 previously, Twitter does allow a user to provide location information, via the `location` field in the user domain. There are also several variables that allow for inference of the user's rough geographic location (e.g. `time_zone`), as well as experimental features that, on the other hand, provide rich metadata on a particular location.

For this thesis, I will be discounting the use of the *Places* feature on Twitter (metadata item `place`) as it is still in the experimental phase. Backdated user/message records will not have valuable information with regards to these features.

On the other hand, I have considered the usage of time zones in determining the rough longitude of the user's current location. In fact, several research studies [Java et al., 2009; Schafer, 2010; Krishnamurthy et al., 2008; Kwak et al., 2010] have used the user profile's time zone information (available as `time_zone` and `utc_offset`) to deduce a user's location. A weakness to this approach is that the user time zone can be inaccurate. This can be as trivial as the wrong time zone being set by a user. Another reason is the intentional change of time zone by users, as seen more recently in the Iran Election controversy where users from around the world changed their Twitter time zone to Tehran as a sign of solidarity [Cheong and Lee, 2010b; Burns and Eltham, 2009].

Based on the availability of devices and software clients capable of using GPS and location data to estimate a user's location and attach it to a given tweet, and the fact that users tend to publish their location (in the `location` profile field), the ability to deduce the user's location — or at the very least the country the current user is in — is beneficial in learning the properties of users.

Two Phase Geolocation Approach

Based on the justifications in the previous subsection, I propose two methods [Cheong and Lee, 2010c] used in conjunction with one another, to determine the country a particular Twitter user is currently residing in.

1. **For tweets with accurate location data:** If the `coordinates` object is present in message metadata, or a latitude/longitude pair is available in the user metadata *location* string, such geographic coordinates can be used to directly and accurately ascertain the country by *reverse-geocoding.*

2. **For tweets without accurate location data (especially tweets collected in the early stages of my research):** The free-form location text presented by the user in the `location` user metadata item is used. In the latter case, however, the location field can be populated by names of places with different levels of detail. Examples can range from a specific street (*Flinders Street, Pok Fu Lam Road*), to entire districts/states (*Morwell, Australia* or *Ibaraki Prefecture*). This is a time-consuming operation to map locations to specific countries, but are nonetheless meaningful in the absence of per-message or per-user geographic coordinate data.

Third-party Commercial Geolocation

Following my proposed approach, my initial exploration of geolocation approaches led me to select the Google Maps Geocoder API to look up geographic coordinates, and query for the country from a user-supplied location string [Cheong and Lee, 2010c]. The rationale is that Google has a comprehensive API which is free, programmer-friendly, and has an extensive set of location names built upon their rich Google Maps service. The approach improves on existing ideas such as the usage of the Yahoo Geocoding API which only worked for GPS coordinates [Java et al., 2009].

To determine the accuracy of using the Google Maps Geocoder API, I conducted Experiment 4.4.

Experiment 4.4. *To measure the accuracy of my proposed two-phase geolocation approach, where both phases are outsourced to the third-party Google Maps Geocoder API.*

METHOD: I ran reverse geocoding (coordinate lookup) and also location string lookup based on the Google Geocoder API on ten similar data sets, totaling 1,000 user records [Cheong and Lee, 2010c]. I will refer to this dataset throughout this thesis as *1k-Locations*.

As ground truth to *1k-Locations*, human volunteers manually identify the locations of the places — with the aid of the Google search engine[8], the OpenStreetMap atlas site, Windows Live Maps (now Bing Maps), and Wikipedia — and categorize them according to countries. For consistency, countries are uniquely identified by their two-character ISO-3166-1 country codes to avoid conflict in naming conventions (e.g. the ISO-3166-1 code of `CD` to clearly refer to the *Democratic Republic of the Congo*, which is also formerly known as *Democratic Republic of Zaire*).

RESULTS AND DISCUSSION: Table 4.5 lists the findings from my validation testing.

Test set	1	2	3	4	5	6	7	8	9	10
Algorithm performance: percentage of correctly determined countries	96	90	87	90	92	92	88	93	80	89

Table 4.5: Location string lookup and reverse geocoding: algorithm accuracy (versus ground truth)

An average of 90.7% detection accuracy was achieved using the Google Geocoder API for my two-phase geolocation approach. The obtained accuracy assumes that the human tester knows exactly where a particular location is in the world, and which country it exactly belongs to. Strong cases would be for GPS coordinates, and the presence of a full address.

Several cases have been identified, however, where the location matching mechanism becomes weak. This include random and nonsensical place names (e.g. `somewhere in the world`, the listing of multiple locations or misspellings of locations (e.g. `London/Paris/Tokyo`)

[8] Excluding the Google Maps service, which is a part of the Geocoder API, to avoid bias.

and `N.Y.Cc`), and the ambiguity of locations (e.g. `Brighton Beach` could refer to different places in both the United Kingdom or Australia).

Proposed Algorithm for Scalable and Robust Offline Geolocation

However, as the work on this thesis evolved, several drawbacks were noticed in the original proposal of using a third-party geolocation service.

I have identified several drawbacks in third-party geolocation and geocoding services such as the Google Geocoding API [Cheong and Lee, 2010c] above or Yahoo Placefinder API [Java et al., 2009] in terms of:

- **Cost:** the cost of using commercial third-party services are prohibitively high if geocoding was to be applied to a very large data set.
- **Licensing restrictions:** compared to the time of writing of [Cheong and Lee, 2010c], Google has since restricted their API to be used strictly for map generation in conjunction with the Google Maps platform[9].
- **Usage quota and volume of data:** as online geocoding significantly consumes computational resources and network bandwidth, it is not ideal for large-scale geocoding of Twitter messages. Furthermore, there are hard limits imposed on the quota of location records that can be geocoded in a window of time. Again, a high price factor is involved when a particular quota is reached.

The above drawbacks prevent this approach being scaled to handle millions of records, especially since the Streaming API has superseded the original on-demand APIs (Section 4.2.3). Prior work, e.g. [Java et al., 2009; Krishnamurthy et al., 2008], mine inclusive [Cheong and Lee, 2010c], dealt only with a magnitude of thousands.

Based on my initial identification of the two phases involved in geocoding Twitter data, I propose a novel *Two-phase Hybrid Geocoding* method using open source and public domain data, that can easily be scalable as required to handle differing amounts of input. My proposed approach involves the following two steps, used in tandem with one another:

1. **Coordinate reverse-geolocation:** If a coordinate point — expressed as a latitude-longitude pair — is found in Twitter message metadata (the `coordinates` extended metadata object) this point is reverse-geolocated to determine the country in which it belongs to. Message-specific metadata is favored, as it is generated every time a Twitter user publishes a tweet with a supported device with geotagging activated. In its absence, the `location` field in user metadata is checked to see if it contains a coordinate point, which is commonly performed by older mobile software as early as 2009 [Cheong and Lee, 2009].

2. **Free-form string parsing:** For records without coordinates in user and message metadata, but with a free-form `location` string in user metadata, I attempt to parse

[9] As described in the Terms of Service from the Google Geocoding API documentation: <`https://developers.google.com/maps/documentation/geocoding/`>.

the string for locations using the open-source *Geodict* algorithm Warden [2011]. Any returned coordinates from *Geodict* can then be parsed using coordinate reverse-geolocation.

Offline coordinate reverse-geolocation technique

I use public domain data on country boundaries from Natural Earth [*Natural Earth*, 2012]. The data is provided in the *ESRI Shapefile* shape format, which stores a series of coordinate points outlining the border of each country, which is represented as a polygon. Shapefile *geospatial metadata* on every country is included as part of the Shapefile and can be accessed to return attributes for each country on the map.

A polygon hit test is used to see if a particular geographic coordinate point (latitude-longitude pair) is located within a given country's polygon. If it is, the country's ISO-3166-1 two-character code, as stored in the geospatial metadata, is returned.

To improve lookup speed, I applied the *Quadtree* algorithm [Finkel and Bentley, 1974] to preload the polygon data (boundary points) in memory. Using a quadtree trades off memory in favor of search speed by narrowing the search space, and is commonly used in algorithms related to cartography and geographic information systems. Algorithm 4.2 illustrates the inner workings of this method, which was published in [Cheong et al., 2012b].

Algorithm 4.2 Offline coordinate reverse-geolocation algorithm for latitude-longitude pairs.

```
procedure InitializeQuadtree(quadtree)
    initialize quadtree by segmenting and adding levels
    load shapefile
    for each country in shapefile do
        centroid ← calculate centroid coordinates
        traverse quadtree to determine where centroid belongs
        add node at centroid tagged by country
    end for
    return quadtree
end procedure
procedure CoordinateReverseGeolocation(coordinates)
Require: quadtree ← initialized by InitializeQuadtree
    traverse quadtree to find coordinates
    searchspace ← regions in quadtree containing coordinates
    for each country in searchspace do
        countryshape ← look up vector shape data in shapefile for country
        if coordinates satisfies polygon hit test for countryshape then
            countrycode ← look up country metadata in shapefile for country
            return countrycode
        end if
    end for
end procedure
```

A consideration that needs to be made is that Natural Earth [*Natural Earth*, 2012] provides vector map data in one of three levels of detail: a scale of 1:10m (most detailed), 1:50m, and 1:110m (least detailed). To determine the suitability of each of the three levels of detail, I conducted Experiment 4.5 as follows.

Experiment 4.5. *To determine the speed and accuracy of reverse geo-location for each detail level of Natural Earth map data, for deciding the optimal map scale for use in Algorithm 4.2.*

Method: A list of 232 countries, their capitals, and the capitals' known geographic coordinates were generated using Wolfram Mathematica's knowledge engine. For each country, its capitals' geographic coordinates were fed into Algorithm 4.2; the algorithm's result is compared against the actual country the coordinates belong to.

Results and Discussion: The difference in accuracy between the different scale levels is caused by the approximation of a country's borders by discrete points, causing locations near the boundary to be wrongly detected as being part of another country or no country at all (e.g. coastal regions). Table 4.6 summarizes the experimental findings, with respect to each of the scale levels mentioned.

Level of detail (scale)	**Computation speed (average)**	**Accuracy**
1:110m	0.3922 sec (∼ 0.0017s per input)	67.24% (156 out of 232)
1:50m	18.6402 sec (∼ 0.0803s per input)	82.76% (192 out of 232)
1:10m	278.7290 sec (∼ 1.2014s per input)	87.07% (202 out of 232)

Table 4.6: Evaluation of computational speed and accuracy for the different scales of Natural Earth map data.

From my evaluation, the choice of 1:100m clearly favors speed, by heavily sacrificing on accuracy. The scale of 1:10m has the highest accuracy, but is very slow at approximately 1.2 seconds per input string, even with the quadtree algorithm [Finkel and Bentley, 1974] heavily narrowing down the search space. This is infeasible for large data processing. Hence, I decided to use the 1:50m scale as it strikes a balance of both accuracy and speed [Cheong et al., 2012b].

Free-form string parsing approach

For records without coordinates in user and message metadata, but with a free-form `location` string in user metadata, I proposed the use of the *Geodict* algorithm [Warden, 2011] to extract location strings from the free-form location text, and map it to real world locations.

Geodict, as part of the *Data Science Toolkit*, is an open-source algorithm by Warden [2011]. It works by extracting tokens from a given string, and attempts to match it with approximately four million records of real-world locations stored in a relational database in order to deduce a particular geographic location (ranging from city level to country level).

The inner workings of *Geodict* are illustrated in Algorithm 4.3 [Warden, 2011].

Algorithm 4.3 Geodict algorithm by Warden [2011], to parse free-form location strings.

```
procedure GEODICT(string)
Require: db $\leftarrow$ contains place names and metadata
    for each word substring in string in reverse order do
                                        $\triangleright$ e.g. “Paris France” becomes “France Paris”
        perform simple string matching of substring against db
        if substring is not a string found in db then
            next
        end if
        for each locationformat do                $\triangleright$ e.g. “Region, Country”,
            for each token of locationformat in reverse order do
                        $\triangleright$ e.g. for Region, Country: country token, before region token
                if tail of substring $\neq$ tail of token from db then
                    break
                else
                    tokenresult $\leftarrow$ matching token
                    expand substring by one word from string
                end if
            end for
            if tokenresult is valid then
                update substring
                break
            end if
        end for
        if tokenresult is not valid after all token testing then
            discard current substring
            substring $\leftarrow$ preceding word from string
        else
            geographicdata $\leftarrow$ read db for geographic data on substring
        end if
    end for
    return geographicdata
end procedure
```

Given a free-form `location` string, *Geodict* can directly consume it (i.e. the input parameter *string* in Algorithm 4.3). If the algorithm successfully finds a match, it returns the parsed location's geographic data. One of two cases will occur:

- If the parsed location is a subdivision of a country (e.g. town or city), a set of coordinates is obtained from *Geodict*'s output which would be parsed using my coordinate reverse-geolocation technique proposed earlier.

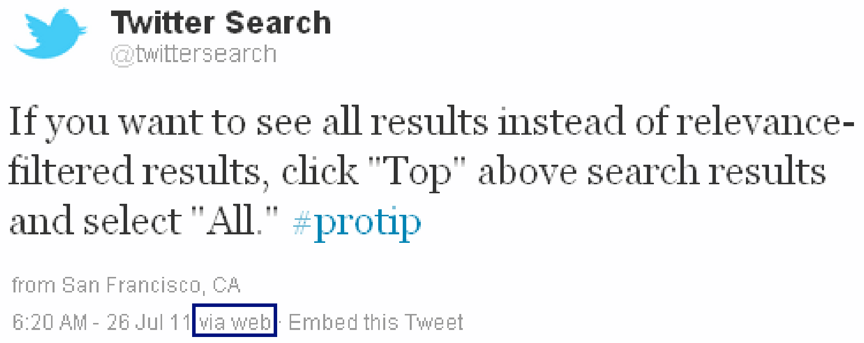
- If the parsed location is a country, *Geodict* also returns the ISO-3166-1 two-character code which can be used directly.

The advantage of *Geodict* and the *Data Science Toolkit* is that it contains open, non-commercial, mapping data that can easily be updated or expanded (using a relational database). This implementation of *Geodict* can easily be scaled up to handle a large number of messages, as a copy of the database is hosted on the *Amazon Elastic Compute Cloud* (EC2) platform; this can be deployed at will at a very low cost (less than $5 USD per day), and the allocated computing resources can 'elastically' be reallocated to cope with the scale of input data.

I trialed my *Two-phase Hybrid Geocoding* algorithm — combining my coordinate reverse-geolocation in Algorithm 4.2 and *Geodict* free-form string parsing as per Algorithm 4.3 — on *10-Gigabyte Dataset*. The results, detailed later in Section 5.4), indicate that my proposed hybrid algorithm performed rather well in terms of feasibility and ability to cope with the volume of data [Cheong et al., 2012b].

4.6.3 `source` Metadata and Device Classes

An important but oft-overlooked item in the message metadata is the presence of the `source` attribute, which stores the name of the program or interface a particular Twitter message was composed with. This is visible to an end-user of Twitter's web interface, as a brief text string at the end of every tweet; such as the string "...via web" highlighted in Figure 4.6.


Twitter Search
@twittersearch

If you want to see all results instead of relevance-filtered results, click "Top" above search results and select "All." #protip

from San Francisco, CA
6:20 AM - 26 Jul 11 via web · Embed this Tweet

Figure 4.6: A sample tweet by user `@twittersearch`, with the `source` metadata item — in this case the *web* interface — highlighted with a blue rectangle.

Proposed Classification Schemes for Device Classes

The methodology behind prior related work involves extracting a sample of source strings from their dataset of Twitter messages and categorizing the `source` string according to the type of software it is [Cheong et al., 2012b]. These studies involve the usage of the old REST APIs for small-scale empirical data collection.

Krishnamurthy et al. [2008], who surveyed the different `source`s in their dataset came up with a list of five categories:

1. *Web*
2. *txt* (mobile)
3. *Instant Messaging* applications
4. the Twitter application for *Facebook*
5. other *custom applications*

Java et al. [2009] came up with three categories, a mere subset of the list by Krishnamurthy et al. [2008]: the Twitter website, SMS, and instant-messaging agents: i.e. the first three items in the list by Krishnamurthy et al. [2008].

My initial small-scale study [Cheong and Lee, 2009], documented in Experiment 4.6 used the `source` metadata item to infer a *device class*, or software category.

Experiment 4.6. *Initial study [Cheong and Lee, 2009] to devise a categorizing scheme for software — or device classes –from* `source` *metadata.*

METHOD: This initial study, conducted in 2009 [Cheong and Lee, 2009], involved 484 Twitter messages from various topics. These messages were collected using the old `search` API as part of a clustering study, which will be discussed in Section 6.2 later. I manually extracted and categorized each `source` string acquired from each of the 484 tweets. These `source` strings were collated, before I manually searched the Internet to find out more information about the specific software named in each `source`.

RESULTS AND DISCUSSION: From this study [Cheong and Lee, 2009], I have arrived at a list of six device classes, as per Table 4.7

A follow-up analysis was performed in my paper on automated metadata analysis [Cheong and Lee, 2010c]. In this experiment (Experiment 4.7), authored in late 2009 and published as [Cheong and Lee, 2010c] in 2010, the data set is significantly larger than the one in Experiment 4.6.

Experiment 4.7. *Follow-up study [Cheong and Lee, 2010c]; to categorize software by device class, by evaluation of* `source` *metadata from a 14,000 message sample.*

Table 4.7: My categorization results from Experiment 4.6 [Cheong and Lee, 2009]

Device class	**Explanation**
web	The main Twitter website at <`http://www.twitter.com`>
mobile	The mobile Twitter website at <`http://m.twitter.com`> or the Twitter SMS interface.
social media	Includes Twitter for Facebook applications and other social media-based Twitter clients.
RSS	Programs which post tweets based on RSS feeds.
marketing	Twitter clients which are mainly used in marketing campaigns.
other	Other Twitter programs.

Method: This experiment involved the categorization of 66 unique software clients in the `source` variable on a case study of approximately 14,000 messages. Similar to Experiment 4.6 [Cheong and Lee, 2009], the `source` strings from each message were extracted and collated into a list. For each unique `source` string, I selected an appropriate device/platform class via the software authors' descriptions in the website/download page of the software; failing which, by conducting a simple web search.

Results and Discussion: The results from this follow-up analysis [Cheong and Lee, 2010c] extended the categorization performed in prior work [Cheong and Lee, 2009; Java et al., 2009; Krishnamurthy et al., 2008] to provide a clearer overview of the various techniques users contribute to the '*Twitterverse*'. I obtained the following list of device/platform classes:

Table 4.8: Follow-up: categorization results from Experiment 4.7 [Cheong and Lee, 2010c]

Device class	**Explanation**
web	The official Twitter web interface.
mobile devices	The mobile Twitter website at <`http://m.twitter.com`>, Twitter SMS interface, and third-party Twitter mobile software.
social media	All social media applications.
alternative clients	Other Twitter client software or interfaces.
feed aggregators/RSS	Programs which post tweets based on RSS feeds.
Twitter 'mash-ups'	Software which 'mashes-up' Twitter data with information sharing.
Twitter marketing tools	Twitter tools which are used for marketing purposes, including bulk messaging tools.
other	Other Twitter programs.

Improved Classification of Device Classes with Large-Scale Empirical Data

The studies mentioned in the prior subsection [Java et al., 2009; Krishnamurthy et al., 2008; Cheong and Lee, 2009, 2010c], despite being good foundations for studying device classification on Twitter `source` strings, have two major limitations [Cheong et al., 2012b], viz.:

1. **The number of category labels for classification**: This is due to the rapid development of new Twitter software, and hence an increase of source strings [Cheong et al., 2012b].

2. **The lack of empirical findings into the various categories of `source` strings**: This is again due to the magnitude of prior research which only covered samples in the order of thousands. [Cheong et al., 2012b].

With the availability of the Twitter Streaming API, and consequently, my collection of the *10-Gigabyte Dataset* (Section 4.5) from said API, I was able to conduct a large-scale classification exercise (Experiment 4.8) to expand upon the proposed definitions [Java et al., 2009; Krishnamurthy et al., 2008; Cheong and Lee, 2009, 2010c] based on real-world empirical data [Cheong et al., 2012b]. This dataset contains raw metadata extracted from 7,863,650 messages; more details of this dataset will be elaborated in Section 5.4.

Experiment 4.8. *To perform large-scale classification — expanding upon Experiments 4.6, 4.7 and [Cheong and Lee, 2009, 2010c]; and also prior device class definitions [Java et al., 2009; Krishnamurthy et al., 2008] — using 7,863,650 real-world data records [Cheong et al., 2012b].*

METHOD: Similar to my methods in both [Cheong and Lee, 2009] and [Cheong and Lee, 2010c] for each of the records in the *10-Gigabyte Dataset*, I collated all the software client strings in the `source` metadata fields found in the *10-Gigabyte Dataset*. This time, I organized them in a frequency distribution beforehand; frequency bins with slight string differences caused by escape characters or artefacts from character encoding were merged (e.g. "`Twitter for BlackBerryÂ®`" [*sic*] and "`Twitter for BlackBerry`"). The resulting frequency distribution is comprised of 29,097 unique software client `source` bins.

From this, I narrowed-down the 300 most frequently-used software clients, made up of 96.8% of `source` strings found in the dataset. A complete listing of software clients is included in Appendix B. By searching the Internet to deduce the type of software a `source` string refers to, I classify each of the 300 `source` strings into suitable device classes using existing findings in Experiments 4.6 and 4.7 [Cheong and Lee, 2009, 2010c] as the seed list. Hitherto undiscovered source strings are collated into new groups based on their similarity.

RESULTS AND DISCUSSION: The device classes found through analysis of empirical data using the methodologies above [Cheong and Lee, 2009, 2010c] yielded a new categorization scheme of 14 device classes. (A complete trend analysis of the data found in the *10-Gigabyte Dataset* will be discussed at length in the following chapter, in Section 5.4).

Table 4.9 enumerates the device classes from this classification exercise (with observed usage frequency in parentheses); sorted in descending order of observation frequency in the *10-Gigabyte Dataset*.

Other software client `source` strings found in the sample — i.e. the 3.20% in the *long tail* of the distribution — are classified as *others.* These `source` strings contain rarely-used

Table 4.9: Categorization results from Experiment 4.8, containing 14 device classes as category labels, obtained by analysis of the large-scale *10-Gigabyte Dataset* [Cheong et al., 2012b]

Device class	**Explanation**
mobile devices	Software allowing Twitter use from a mobile platform, including the mobile Twitter website <`http://m.twitter.com/`>, Twitter's SMS interface, and both official and third-party Twitter software for mobile phones (47.27%).
web	the official Twitter website at <`http://www.twitter.com/`> (32.15%).
social network integration	'Apps' found in online social networks (OSNs) such as Facebook which allow Twitter use from within the other OSN (4.20%).
Web 2.0 integration and sharing	Web 2.0 or social media services that integrate with Twitter for the purpose of content-sharing (3.72%).
interfaces (third-party)	Third-party Twitter programs or interfaces to Twitter, usually with many advanced features compared to the official software (2.51%).
feed aggregators	Web services that generate tweets from other feeds, such as Really Simple Syndication (RSS) feeds (2.12%).
bots	Automated or artificial intelligence-based bot programs that publish tweets. From the dataset, it is observed that such software is a niche to the Japanese Twitter community, as five out of six identified bot programs had Japanese websites catering to the Japanese market (2.11%).
marketing tools	Programs or web services used for marketing purposes, such as bulk messaging tools and group-based automated Twitter software, which includes applications exhibiting spam-like behavior (1.17%).
alternate proxies	Mostly intended for mobile devices, these sites are Web based 'proxies' allowing people to access Twitter (0.54%).
Twitter-based third-party sites	Software or web services which provide a novel service (such as fancy visualization) which piggybacks on Twitter as the underlying technology (0.52%).
branded	'Branded' programs or software clients which promotes a particular brand, celebrity, or organization (0.13%).
games	Games utilizing, or promoted, via the Twitter platform (0.11%).
access gateways	Services which allow Twitter use over other protocols, such as email, game consoles, and Internet Relay Chat (0.05%).
suspicious	Programs, websites, or extensions which are dubious in nature, where the original website that hosts the software is suddenly no longer available, or is suspected of rogue behavior such as undesirable browser toolbars (0.19%).

software clients, which have a total observation frequency that (less than 323 out of over 7.8 million) is rather low in the *10-Gigabyte Dataset*. Examples include Twitter clients ‘branded’ and integrated with existing popular websites (e.g. the news site *The Huffington Post*), custom-built bespoke Twitter bots (e.g. those created via a scripting language for experimental purposes), and spam applications.

Using Classifications Programatically as Training Data

Using the classification results in Table 4.9, one can then trivially implement a categorization algorithm to process new Twitter data. Each of the software `source` strings studied can be stored as keys in a hash table, with its corresponding device class in the bucket. Using the first few popular `source` strings that I obtained in Experiment 4.8, a hash table can be constructed with the key-value pairs listed in Table 4.10, with strings normalized to lowercase.

Table 4.10: Sample text dump from a hash table containing the first ten key-value pairs of `source` strings and their corresponding device classes, obtained as a result of Experiment 4.8 .

```
{
    ‘‘web’’: ‘‘web’’
    ‘‘twitter for iphone’’: ‘‘mobile’’
    ‘‘twitter for blackberry’’: ‘‘mobile’’
    ‘‘twitter for android’’: ‘‘mobile’’
    ‘‘ubersocial for blackberry’’: ‘‘mobile’’
    ‘‘mobile web’’: ‘‘mobile’’
    ‘‘tweetdeck’’: ‘‘social_media_integration’’
    ‘‘echofon’’: ‘‘mobile’’
    ‘‘twittbot.net’’: ‘‘bot’’
    ‘‘keitai web’’: ‘‘web’’

    ...
}
```

Hence, given a source string from a new (unprocessed) user record, a simple hash lookup with a `source` string (as the key into the hash) will suffice in determining the device class for a new record. As per an earlier discussion on hash tables (in Section 4.6.1), hash tables trade off storage space in return for near-constant lookup times, making it scalable for large amounts of input data.

Device Class: Inferring Mobility and Usage Behavior

Based on the categorization scheme above, I can further infer the state of *user mobility*, and postulate *usage behavior.* In this case, message metadata is used to infer properties of their authors.

In *New New Media*, Levinson [2009] stated that Twitter is “intrinsically mobile” in the sense that “the capacity to tweet from [mobile devices]... is a defining characteristic” [Levinson, 2009] of Twitter. Mobile media, such as Twitter, has allowed for communication even in “useless physical places” which are originally “useless for communication” [Levinson, 2009]. He further elaborates that mobile technologies (from the point of view the current discussion: the mobile usage of Twitter) has “liberated [the user] from the home or office [and]... this freedom moved us outdoors” [Levinson, 2009].

With this idea in mind, I suggest the usage of the device class property, proposed above, in inferring *user mobility*: i.e. whether a user is currently at a *fixed* location, or in a *mobile* state. The use of desktop clients, the web interface, or Twitter interfaces only available on a non-mobile computer can suggest a *fixed* user mobility. The usage of mobile clients, the mobile Twitter site or the official SMS interface can suggest the user is in a *mobile* state. Although not perfect (as one can, say, even use a mobile Twitter interface at home), this idea does come in to play when investigating cases of Twitter in crisis and convergence events (e.g. [Hughes and Palen, 2009]).

In my paper [Cheong and Lee, 2011] discussing Twitter usage in civilian response to terror events (further elaborated in Section 7.2), I have applied this hypothesis to reveal user mobility in terror events via Twitter client information (the `source` metadata item). I reinforce this proposal based on Dearman et al. [2008] who observed that users tend to share “time-critical information... in the ‘mobile’ state.” [Dearman et al., 2008]. Hence, I have discovered that “breaking news on Twitter can be attributed to the usage of *mobile devices or social media [applications]...* as civilians would be using it on the move to broadcast the situation or their current feelings/sentiments” [Cheong and Lee, 2011].

Similar to deducing *user mobility*, one can use the categorization scheme of Twitter device classes to determine *usage behavior*. Examples of such inferences on usage behavior are:

1. The usage of Twitter clients in the *social network integration* category suggest that a section of Twitter users also **participate in other Web 2.0 social networking platforms**. Example of such clients would be the official *Twitter for Facebook* application which automatically publishes a new tweet as the user updates his/her status on Facebook; and *Seesmic Desktop* which supports simultaneous publishing of messages/status updates to Twitter, Facebook and Google [Cheong and Lee, 2009].

2. The usage of Twitter content-generating tools in the *feed aggregators* device/platform class indicates that a Twitter user is merely **republishing existing content**, e.g. from RSS feeds, which might in turn be syndicated from blogs [Cheong and Lee, 2009].

3. *Marketing tools*, on the other hand, allow **effective information broadcasting or advertising via tweets**, and make it easier for marketing staff to communicate with customers via a brand’s Twitter account. Frequent appearances of this `source` string in a user’s tweets can suggest the fact that the account belongs to a corporation or organization.

I have also identified several potential applications of this classification [Cheong and Lee, 2010c], which include:

- **Detecting censorship:** As seen in recent events e.g. the 2009 Iran Election controversy [Fleishman, 2009], there were alleged attempts to censor Twitter by governments, especially by trying to cut off communication networks. The disproportionate absence of, say, mobile clients in Twitter metadata collected during such events can be used as an indicator of censorship.

- **Determining the extent of mobile and ubiquitous computing amongst the Twitter user base:** Social science researchers, and market researchers can study the shift in microblogging patterns among a particular section of the Twitter user base (e.g. geographically, or by tweet keywords). This can then be used in, say, targeted advertising, or for policy-making.

4.7 Learning by Inference: Online Presence

4.7.1 Profile Customization

Twitter users are allowed to customize their profile with different profile pictures, background images, and color schemes. Several profile items that can be customized — as described in Section 4.3.2 — include a profile picture, background image or color, background image style, text color, link color, sidebar border and fill color. These metadata items in the user domain are used by Twitter internally to render the user's profile page on the Twitter website.

Profile Customization and the User

Given such information, I postulate that the *degree of profile customization* exhibited by a user, when used in conjunction with clustering and pattern recognition algorithms (to be discussed in Chapter 6), can reveal several traits about the user.

In a study by Nowak and Rauh [2005], the authors note that the presence of an avatar (or *profile picture* in the Twitter context) via a computer-mediated medium, will help in "identifying, recognizing, and evaluating [other users]... in the mediated world of geographically distant communication". This sentiment was also shared by Erickson [2008], in his discussion on user visibility. Besides the profile picture, other forms of profile customization such as background and color scheme changes demonstrate individual personalization of a user's Twitter experience. Such exhibitions of profile customization are evidence that Twitter users who customize their profile aim to reflect online presence, and are more likely to interact and participate in Twitter activity as opposed to those who do not [Cheong and Lee, 2010c]. In a study by Schafer [2010], non-spammer users are found to actively customize and complete their profile on Twitter, compared to spam accounts.

Conversely, spammers are more inclined to use the generic settings created upon account registration [McFedries, 2009; O'Reilly and Milstein, 2009], and lack such profile customization. A simple experiment (Experiment 4.9) is devised, wherein I prove the notion that spam Twitter accounts created using automated tools alluded to in Table 4.9 lack customization compared to *de facto* Twitter accounts.

Experiment 4.9. *To check for existence of (or absence of) profile customization in accounts generated by automated bot programs.*

Method: I performed a survey of the existing features of Twitter bot programs which automate the process of creating new Twitter accounts and broadcasting tweets. Using Google to perform a search on the query "`automated twitter account creator`" (via the URL <`http://www.google.com/search?q=automated+twitter+account+creator`>), I surveyed the first few results returned, which contain links to said bot programs. The features and intended effects of the bot programs are then documented.

Results and Discussion: By checking the feature overview and user interface screenshots of such programs, this experiment revealed that current versions of these bot programs (at time of writing) are only capable of providing custom names, passwords, and email addresses for Twitter account registration. Profile customizations, such as profile pictures, and profile style customization, are absent.

In the middle of the spectrum, Twitter accounts which are created for marketing but not automated (e.g. maintained by a human marketer) are said to minimally customize their Twitter profile [Collins, 2009]. In Collins [2009], the author surmises that marketers tend to "...set up a Twitter account, *tile a photo of their product as a background design*, *follow* as many people as possible and then sit back and watch the sales graph climb" [Collins, 2009] (emphasis mine). On the opposite side of the spectrum, Thomas et al. [2011], based on Motoyama et al. [2011], have examined advertisements for programs "...[specializing] in the sale of Twitter accounts... including *xgcmedia.com* and *backlinkvault.com...* [where one can] create accounts with *custom profile images* and descriptions" [Thomas et al., 2011], emphasis mine. Based on such findings, it is hence important to keep track of not only uncustomized profiles, but also partially-customized ones.

Quantifying Customization on Twitter

From the metadata readily available in the user domain, I propose a metric, the *average degree of user customization*. This metric is an integer between *0* and *2* (inclusive) and is calculated as follows:

$$\textit{Degree of user customization} = \textit{Avatar presence} + \textit{Profile style customization} \qquad (4.1)$$

The two variables, *avatar presence* and *profile style customization* are assigned values such that:

- **Avatar presence:** If a custom profile image or avatar [Nowak and Rauh, 2005] is used, the avatar presence will be given a value of *1* (*0* otherwise). This is easily determined by checking the Twitter-generated summary variable `default_profile_image`, which equals `false` if a custom profile image is present; this is also reflected by a custom image URL in the `profile_image_url` field.

- **Profile style customization:** If the user profile's style has been customized, profile style customization will be given a value of *1* (*0* otherwise). The Twitter API generates a summary variable `default_profile`: a value of `false` indicating the presence of profile style customization.

I have performed a comprehensive study on the distribution of profile customization scores on the large *10-Gigabyte Dataset* is explained in the following chapter (within Section 5.5). For now, it suffices for me to present empirical and theoretical findings on Twitter user accounts with a degree of customization of *0*.

These accounts frequently exhibit characteristics of spammers [Cheong and Lee, 2009], and will sometimes be detected by Twitter Inc. and banned for violating the terms of use [Cheong and Lee, 2010b]. The banning of such users was hitherto explained in the discussion of gathering user profiles via the REST-`user` API (Section 4.1.3) and work by e.g. Thomas et al. [2011] on spam detection (Section 3.3.4). Such users exhibit behavior such as:

1. **Aggressive *following* habits:** Such behavior is "...not commonly found among normal Twitter users" [Cheong and Ray, 2011]. Such Twitter accounts which tend to "...[*follow*] thousands of people" for marketing purposes are liable to be banned for spam [Collins, 2009].
2. **Promoting bad links:** These spammers broadcast "links to phishing and malware sites and unsolicited advertisements" [Cheong and Ray, 2011]; modus operandi include sending "...off-topic [`@user`-based] replies just to send out their URL" [Collins, 2009].
3. **Erratic behavior:** These spam accounts, as discovered by [Metaxas and Mustafaraj, 2010; Lee et al., 2010], tend to be bogus users who exhibit tendencies such as duplicating tweets, have an erratic follower/friend social graph, and nonsensical/'canned' tweets.

As explained earlier, these accounts are generated *en masse* using scripts or programs that can create a large number of user accounts; the caveat is that personalization (such as avatar or profile customizations) are almost non-existent with such accounts.

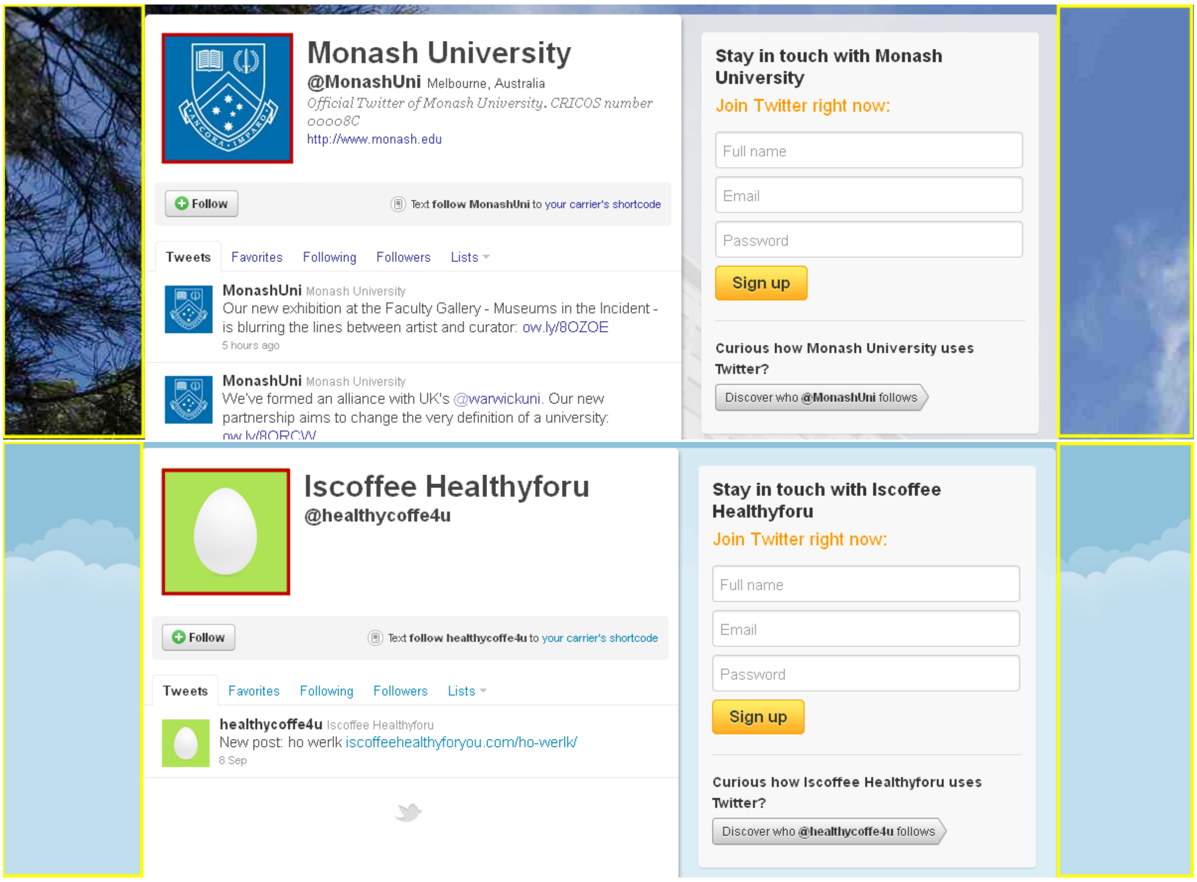
A Case Study: Profile Customization in Reflecting User Characteristics

To illustrate the different degrees of customization, and an example of how this can hint on the characteristics of a Twitter user and reveal spam users, I will provide two example of Twitter user profile screenshots: one being a legitimate Twitter user, and the other one from a Twitter account suspected of spam.

Figure 4.7(a) shows the profile of user (`@MonashUni`), while Figure 4.7(b) illustrates the profile page of user `@healthycoffe4u` [*sic*]. Notice that `@MonashUni` is a legitimate

Twitter account managed by Monash University, with a customized avatar (highlighted in red) and profile style (highlighted in yellow): an average degree of user customization of `2`.

Contrast this with `@healthycoffe4u`, without any avatar customization (highlighted in red) nor profile customization (highlighted in yellow): average degree of user customization of `0`. The latter account exhibits suspicious spam-like behavior with only one tweet composed, and exhibits aggressive *following* patterns with little followers in return (2,001 friends, to 154 followers). The one and only tweet ever composed by this account has a hyperlink that leads to a hacked/rogue website.



Monash University
@MonashUni Melbourne, Australia
Official Twitter of Monash University. CRICOS number 00008C
http://www.monash.edu

Follow
Text **follow MonashUni** to your carrier's shortcode

Tweets | Favorites | Following | Followers | Lists

MonashUni Monash University
Our new exhibition at the Faculty Gallery - Museums in the Incident - is blurring the lines between artist and curator: ow.ly/8OZOE
5 hours ago

MonashUni Monash University
We've formed an alliance with UK's @warwickuni. Our new partnership aims to change the very definition of a university: ow.ly/8ORCW

Stay in touch with Monash University
Join Twitter right now:
Full name
Email
Password
Sign up

Curious how Monash University uses Twitter?
Discover who **@MonashUni** follows

Iscoffee Healthyforu
@healthycoffe4u

Follow
Text **follow healthycoffe4u** to your carrier's shortcode

Tweets | Favorites | Following | Followers | Lists

healthycoffe4u Iscoffee Healthyforu
New post: ho werlk iscoffeehealthyforyou.com/ho-werlk/
8 Sep

Stay in touch with Iscoffee Healthyforu
Join Twitter right now:
Full name
Email
Password
Sign up

Curious how Iscoffee Healthyforu uses Twitter?
Discover who **@healthycoffe4u** follows

Figure 4.7: Screenshot of the user profiles of: (a, above) `@MonashUni`; and (b, below) `@healthycoffe4u`. Note the difference within the profiles, in terms of profile picture/avatar presence (denoted by red squares), and profile background customization (denoted by yellow rectangles).

Inference on Average Degrees of User Customization

From the observations in the previous subsection — how profile customization can allude to the behavior for a particular Twitter user — I decided to extract statistics on average degree of profile customization amongst users who publish tweets highly related to spam and unsolicited messages (Experiment 4.10).

Experiment 4.10. *To evaluate the average degree of profile customization for users tweeting about a spam topic (with a non-spam topic for control).*

METHOD: For this experiment, I will use the *10-Gigabyte Dataset* as the source of Twitter metadata. From the *10-Gigabyte Dataset*, I scan through every message to look for mentions of a common spam phrase, `make money`. User metadata belonging to authors of those tweets are checked for profile customization.

As a control, tweets with the benign greeting "*good morning*" are extracted, and their user metadata on profile customization studied. For comparative purposes, baseline summary statistics for the entire *10-Gigabyte Dataset* are also included.

RESULTS AND DISCUSSION: Table 4.11 illustrates the statistics from this experimental study. From the summary table, even though the size of the test sets are approximately the same, the set of user metadata obtained from tweets containing the spam term "*make money*" has approximately eight times the number of uncustomized profiles (i.e. profile customization = zero) compared to the control set.

This experiment illustrates that the proposed quantization of user profile customization from Twitter user metadata could be used to detect users (and by extension, their messages) with a disproportionate amount of spam; and are likely to violate Twitter's terms of service resulting in account banning. This supplements current studies of spam pattern detection, which currently gauge the user's social links on Twitter to check for possible spam e.g. [Moh and Murmann, 2010; Lee et al., 2010; Thomas et al., 2011].

Message set	**Users with cust. score = 0**	**Users with cust. score = 1**	**Users with cust. score = 2**	**Average customization score (for set)**
"`make money`" (size = 2,489 records)	8.18%	15.19%	76.63%	mean $\mu = 1.6845$ s.d. $\sigma = 0.6161$
"`good morning`" (size = 2,825 records)	1.17%	15.36%	83.47%	mean $\mu = 1.8230$ s.d. $\sigma = 0.4111$
Baseline, *10-Gigabyte Dataset* (size = 2,500 records)	0.29%	24.04%	75.67%	mean $\mu = 1.7538$ s.d. $\sigma = 0.4375$

Table 4.11: Comparison of summary statistics on profile customization, for suspected spam and non-spam (control) terms.

4.7.2 Web Presence

Some users include a profile URL such as a homepage or a blog that will appear when their profile is viewed by others. It is also accessible via the Twitter API in the form of the `url` metadata field. Not to be confused with the appearances of URLs in tweets (i.e. the message domain) however, the `url` metadata item belongs to the user domain. A given user profile can only store one URL.

However, academic research on Twitter rarely focuses on this metadata item. One such study by [Cormode et al., 2010] (discussed prior in Section 4.3.2) reported that as the Twitter user base evolves over time, users tend to use URLs of celebrity fan pages as opposed to URLs of their own pages (such as personal websites or blogs)[10].

Initial Research: Proposed Classification Scheme for User URLs

In my preliminary studies [Cheong and Lee, 2009, 2010c] on a small-scale data set due to the limitations of the REST-`user` API, I observe that a number of users advertise URLs to their profiles on other social networks (such as Facebook), blogs (such as Wordpress), or sites in which they share their content (such as YouTube, and Flickr) [Cheong and Lee, 2010c].

Hence, my initial experiments on the `url` metadata item aims to filter out and categorize the users' URLs based on patterns or stereotypes observed based on the kind of URL published. I opine that this attribute is useful to determine any connections between the users and their corresponding online persona, social media usage, and information-sharing properties [Cheong and Lee, 2010c]. This, in turn, is based on earlier studies on social sharing behavior and information disclosure in online communities [Dearman et al., 2008; Schrammel et al., 2008].

Initially, I observed five main 'stereotypes' of websites that are found in `url` metadata strings:

1. blogs
2. media sharing sites
3. other microblogs (or Twitter clones)
4. Facebook (an OSN)
5. MySpace (another competing OSN)

Using these preliminary empirical observations, I came up with a simple classification method for Twitter profile URLs based on domain name.

A domain name consists of a server or host name, followed by a three-letter top-level domain (TLD), which may also be coupled with a two-letter country TLD (as per ISO-3166-2, seen before in Section 4.6.2). For example, `facebook.com` is the domain name extracted from the `url` string "`http://www.facebook.com/username`").

Using rule-based matching, I assign categories to the URLs based on the five stereotypes listed above [Cheong and Lee, 2009]. If the URL does not match any of the stereotypes above, four generic categories are used to cluster the URLs by their TLD. The list of eight categories used in my categorization scheme, published in [Cheong and Lee, 2010c] is as per Table 4.12.

My initial research involving stereotypes enabled me to determine connections between a group of users and their corresponding social media usage and information-sharing properties.

[10] A version of this phenomenon is also documented by [Levinson, 2009] as a form of roleplaying, where users set their Twitter usernames to fictional characters, e.g. Don Draper from the *Mad Men* TV series.

Table 4.12: Initially-proposed Twitter profile URL categorization scheme with eight labels/stereotypes [Cheong and Lee, 2010c].

Category / Stereotype	**Pattern**
The URLs are first checked against these specific domains in (1–5) to identify website stereotypes.	
1. *Blogs*	`WordPress.com`, `Xanga.com`, `LiveJournal.com`, `spaces.live.com` (Windows Live Spaces)
2. *Media sharing*	`YouTube.com`, `Flickr.com`
3. *Other microblog services*	`Jaiku.com`, `Pownce.com`, `Tumblr.com`, `Plurk.com`
4. *The Facebook social network*	`Facebook.com`
5. *The MySpace social network*	`MySpace.com`
If the URL does not match the stereotypes (1–5) above, they are matched to a generic category (6–9).	
6. *Educational website*	URLs with the `.edu` TLD
7. *Organizational website*	URLs with the `.org` TLD
8. *Personal/commercial website*	URLs with the `.com` or `.net` TLDs

An example can be found in a case study I conducted in [Cheong and Lee, 2010c] and later expanded in [Cheong and Lee, 2010b]. Among the users discussing the 2009 Iran Election controversy [Fleishman, 2009; Ems, 2010], I was able to deduce a group of clustered 'veteran' Twitter participants who also record a "...high usage [activity] of other social media sites such as owning a blog or social network page" [Cheong and Lee, 2010c], which illustrates Twitter's usage by bloggers as a complement to their traditional blog posts.

Distribution of Profile URLs from Large-Scale Empirical Observations

My creation of the *10-Gigabyte Dataset* (Section 4.5) allowed me to conduct a large-scale classification exercise for profile URLs in Twitter user metadata. This is akin to the device classification exercise in Section 4.6.3, and is documented in Experiement 4.11. Complete analysis of the empirical data found in the *10-Gigabyte Dataset* is located in Section 5.5.

Experiment 4.11. *To extract profile URLs from user metadata in the 10-Gigabyte Dataset, which are then classified and categorized by domain name.*

METHOD: Based on [Cheong and Lee, 2010c], I extract domain names found in the `url` string found in user metadata, and collated them in a frequency distribution. A total of 1,536,729 unique user records containing URL strings were recorded. From this, I have isolated 338,655 unique domain names. As per the analyses conducted on device classes (Section 4.6.3), the 300 most frequently-occurring unique domain names, found in approximately 72% of the 1,536,729 URLs found, are taken. By visiting the website pointed to each of the domains, I devise a classification scheme for these 300 domains, with my originally proposed classification scheme [Cheong and Lee, 2010c] as the seed list.

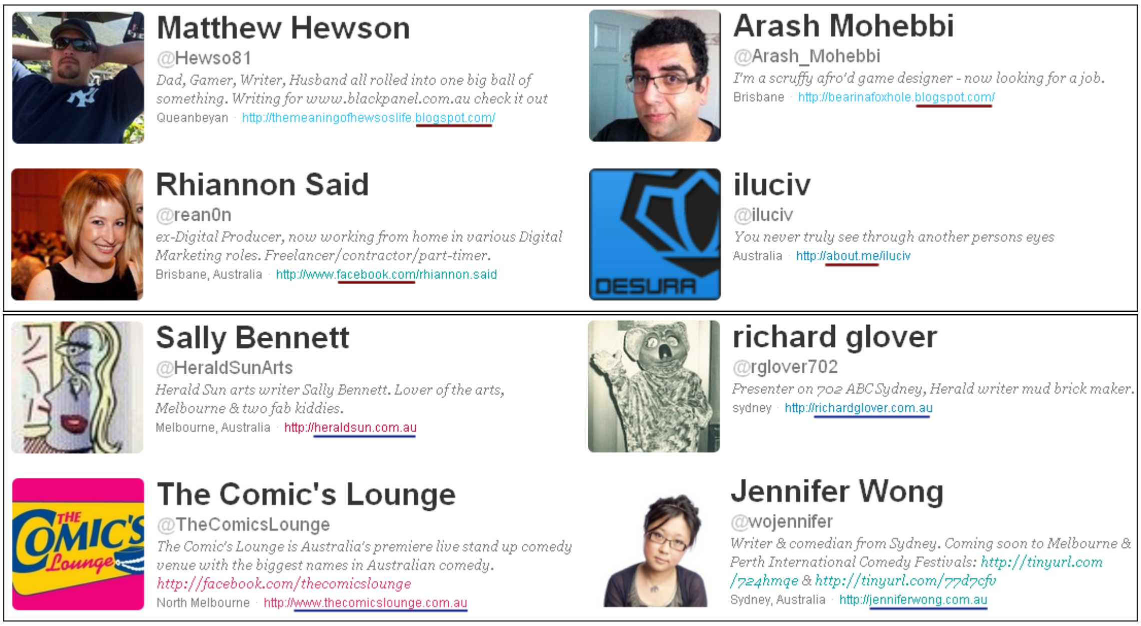
The raw data table resulting from this classification exercise is provided for reference in Appendix B.

Results and Discussion: The improved classification scheme for user `urls` based on domain names is enumerated in Table 4.13, in descending order of observed frequency.

The remaining sites which do not belong into any of the above categories are classified as *others*. In the *10-Gigabyte Dataset*, there are 338,355 such sites, which comprise the long tail of the distribution of URL domains. (From the long tail in the *10-Gigabyte Dataset*, each of those such sites comprise at most 91 observations among the 4,491,022 user profiles).

Notes on URL Classification

From existing literature [Krishnamurthy et al., 2008; Cheong and Lee, 2010c], I find that the URL strings in Twitter user profiles are useful in determining connections between a stratum of users and their corresponding social media usage, location, and information-sharing properties. One of the studies with this respect would be Krishnamurthy et al. [2008], who extracted domain information from the URL string, and used the domain names in conjunction with timezone data, to "...see [the] popularity of Twitter in different countries." [Krishnamurthy et al., 2008].



Matthew Hewson
@Hewso81
Dad, Gamer, Writer, Husband all rolled into one big ball of something. Writing for www.blackpanel.com.au check it out
Queanbeyan · http://themeaningofhewsoslife.blogspot.com/

Arash Mohebbi
@Arash_Mohebbi
I'm a scruffy afro'd game designer - now looking for a job.
Brisbane · http://bearinafoxhole.blogspot.com/

Rhiannon Said
@rean0n
ex-Digital Producer, now working from home in various Digital Marketing roles. Freelancer/contractor/part-timer.
Brisbane, Australia · http://www.facebook.com/rhiannon.said

iluciv
@iluciv
You never truly see through another persons eyes
Australia · http://about.me/iluciv

Sally Bennett
@HeraldSunArts
Herald Sun arts writer Sally Bennett. Lover of the arts, Melbourne & two fab kiddies.
Melbourne, Australia · http://heraldsun.com.au

richard glover
@rglover702
Presenter on 702 ABC Sydney, Herald writer mud brick maker.
sydney · http://richardglover.com.au

The Comic's Lounge
@TheComicsLounge
The Comic's Lounge is Australia's premiere live stand up comedy venue with the biggest names in Australian comedy. http://facebook.com/thecomicslounge
North Melbourne · http://www.thecomicslounge.com.au

Jennifer Wong
@wojennifer
Writer & comedian from Sydney. Coming soon to Melbourne & Perth International Comedy Festivals: http://tinyurl.com/724hmqe & http://tinyurl.com/77d7cfv
Sydney, Australia · http://jenniferwong.com.au

Figure 4.8: Screenshots of user profiles including URLs of users who have authored tweets about: (a, above) "*Mana Bar*"; and (b, below) "*Melbourne Comedy Festival*". The profile URLs are underlined in different colors (red for "*Mana Bar*", and blue for "*Melbourne Comedy Festival*"), to emphasize the distinctive URL categories found within each set of users.

I will now briefly illustrate a qualitative evaluation in which user behavior can be hinted upon by virtue of URL categories on a user's Twitter profile. By randomly selecting users on Twitter who authored tweets on two different subjects, and investigating the type

Table 4.13: URL classification scheme resulting from observations in Experiment 4.11, conducted on metadata in the *10-Gigabyte Dataset*.

URL Category	**Description**
Online social networks (OSNs)	URLs to a Twitter user's profile pages on social networks such as Facebook and MySpace. (26.76%)
Other microblogging sites	Profiles on other microblogging sites such as Tumblr. (14.55%)
Blog or personal webpages	These include personal blog sites such as Blogspot and Livejournal, and pages on personal website hosting services. (13.69%)
Media-sharing sites	Profile or gallery pages on media-sharing sites, such as YouTube and Flickr. (10.20%)
Official Twitter pages	Users on Twitter link back to their own Twitter user page (self-referential) or with the intent of promoting another user's Twitter page (2.54%).
Twitter-based media	Twitter users link to other Twitter-related web services, which usually integrates with the Twitter infrastructure or API. An example would be TwitPic, a popular [Terdiman, 2009] photo-sharing service which integrates with Twitter; due to this key property, it is not classified as a *media-sharing site*. (1.62%)
Web portals	Pages on Web portals, such as Google (personalized Google homepages), and Naver (a popular South Korean web portal). (0.98%)
URL shortening services	Some of the links in the dataset have been shortened with URL-shortening services, such as TinyURL, `t.co` (Twitter's URL shortener) and `fb.me` (Facebook's URL shortener). (0.78%)
Twitter user indices / directories	Such sites are listings of popular Twitter users, akin to a 'phonebook' service. (0.19%)
Informational / news sites	Examples of these sites include Wikipedia and IMDB, the Internet Movie Database. (0.32%)
Branding	Some users link to pages featuring their favorite brand, product, celebrity, games or services. (0.24%)
Sales	These comprise of product pages on online e-commerce or retailer websites, such as Etsy and Amazon. (0.16%)
Adult	Sites of an adult nature. (0.03%)
Bots / marketing	These sites feature bots or social marketing tools which help in product marketing by generating automated tweets. I opine that these bot programs are part of aggressive social media marketing campaigns. (0.01%)
Suspicious	Several sites are observed to be suspicious in nature. These websites are inaccessible ("site not found" errors are encountered), and a Google search performed on the website reveals negative feedback by other users. (0.01%)

of URLs they mention in their profile metadata, I am able to infer other characteristics common among like-users.

- "*Mana Bar*": this refers to an Australian video-game-themed bar, whose target audience is young adults in their 20s and 30s who are interested in video games, and popular culture. By observing the type of URLs published by users who tweet on this subject (underlined in red, in Figure 4.8), one can observe that these users frequent online social networks and maintain blogs, which would correlate with the listed target audience (cf. [Bozkir et al., 2010; Argamon et al., 2007]).
- "*Melbourne Comedy Festival*": this refers to an annual comedy festival held in Melbourne, Australia. The users discussing this topic in Figure 4.8 tend to publish URLs which link to Australian informational and personal sites with `.au` domains (underlined in red, in Figure 4.8). This shows that a proportion of tweets related to this topic belong to Twitter accounts which are e.g. newspapers, journalists, or affiliates of the comedy festival. Contrast this with the first case, "*Mana Bar*", where tweets mainly come from young people who fit the target demographic.

The findings from [Dearman et al., 2008; Hughes and Palen, 2009; Schrammel et al., 2008] apply here, as the Twitter profile URL is a publicly available hint of a user's identity; allowing one to gleam an insight into a given Twitter user's online persona.

4.7.3 User Connectivity

Defining User Connectivity in the Twitter Social Graph

Despite not being a full-fledged online social network, Twitter is still categorized as a micro-OSN [Krishnamurthy, 2009], as it still contains characteristics of a social network that allows users to build links amongst one another. On Twitter, each user is asymmetrically connected to others based on the *friend/following* mechanism (Section 2.1) where a user could *follow* another (subscribe to the other user's Twitter updates) without necessarily being *friended* in return.

Studies on user connectivity on Twitter [Java et al., 2009; Krishnamurthy et al., 2008] range from dynamics of social networking and interaction for a particular user or group of users by crawling the social graph via the REST-`user` API, to topological analysis by Kwak et al. [2010] and Huberman et al. [2008a].

The FFR as a Summary Statistic

In analyzing a Twitter user's social connections, [Java et al., 2009; Krishnamurthy et al., 2008] introduced the *follower/friend ratio*; a summary statistic reflecting the connectivity of a user with respect to other users. The FFR is simply the ratio of a user's in-degree (number of people *following* said user) to out-degree (number of people — 'friends' in Twitter API terminology — that said user *follows*) with respect to the Twitter social graph.

With respect to research, this metric is easily obtained via Twitter APIs (both the old REST API and the current Streaming API) which provide the two metadata items `followers_count` and `friends_count`. The FFR is defined[Cheong and Lee, 2010c] as:

$$\text{FFR} = \frac{\text{number of other users } \textit{following} \text{ the current } (\texttt{followers\_count})}{\text{number of `friends' the current user is } \textit{following} \ (\texttt{friends\_count})} \tag{4.2}$$

For cases where `friends_count` equal zero, the FFR be invalid as it equates to infinity due to division by zero. Hence, for the purposes of FFR calculation, `friends_count` values need to be positive integers. To account for this, any `friends_count` of zero will be normalized to a value of one.

Suitability of the FFR in Studying User Behavior

With the Twitter Streaming API superseding the REST API, and the imposition of rate limits to the REST-`user` API (as per Section 4.1.3), the ability to enumerate the complete lists of friends and followers for a given Twitter user is severely restricted. This means that experiments in crawling the user network, construction of a user's social graph, or identification of properties of a given user's friends and followers, e.g. [Java et al., 2009; Kwak et al., 2010] are no longer feasible.

However, as the Streaming API provides both user and message metadata, I was still able to obtain the metadata items `followers_count` and `friends_count`, which are used for the derivation of a given user's FFR. Hence, in this thesis, I use the FFR to summarize a Twitter user's social connections.

The FFR was tested in existing research [Krishnamurthy et al., 2008] to determine the popularity and social networking habits of a Twitter user, simply by the fact that it can directly be used to identify three categories of users:

1. **miscreants** (spammers or stalkers) or *evangelists* who have a low FFR; as they "...contact everyone they can, and hope that some will *follow* them" [Krishnamurthy et al., 2008; Ostrow, 2008], italics mine.

2. **broadcasters** are users which have a high FFR, which "...characterizes broadcasters of tweets [e.g.] online radio stations, who utilize Twitter to broadcast the current song they are playing... [and] other media outlets generating headlines" [Krishnamurthy et al., 2008]. This definition is expanded to include opinion leaders or famous celebrities on Twitter, as similar FFR values have been observed among such popular users [Cheong, 2009].

3. **ordinary users**, termed *acquaintances* by Krishnamurthy et al. [2008], who have an FFR close to 1.0; i.e. a near-equal amount of friends and followers. Krishnamurthy et al. [2008] opines that these are everyday Twitter users who regard Twitter as a social network as they "...tend to exhibit reciprocity in their relationships" [Krishnamurthy et al., 2008].

The FFR is also incorporated into several studies on the use of summary statistics for Twitter spam identification [Moh and Murmann, 2010; Abrol and Khan, 2010; Lee et al., 2010]. In these studies, the FFR summarizes the likelihood of a user being a spammer. In terms of the previous categorization by Krishnamurthy et al. [2008], these are *miscreants*, who tend to have a very low FFR. Abrol and Khan [2010] also adapted the FFR as part of a formula to determine probability of spam for a particular Twitter user. In empirical spam and anomaly analysis studies, the disproportionately low FFRs (high delta between friend and follower counts) amongst spammers or miscreants are characteristic in singling out spam users [Thomas et al., 2011; Schafer, 2010; Barracuda Networks, Inc., 2010].

Case Studies of the FFR in the Real World

To conclude the discussion of the usage of the FFR, I will highlight some anomalous FFR examples from the *10-Gigabyte Dataset* to illustrate the differences between the *miscreants* and *broadcasters*, cf. Krishnamurthy et al. [2008], in Experiment 4.12. A complete evaluation of the FFR distribution for the entire *10-Gigabyte Dataset* follows in the next chapter (Section 5.5).

Experiment 4.12. *To evaluate the characteristics for users with the highest and lowest five FFRs from the 10-Gigabyte Dataset.*

METHOD: All user records in the *10-Gigabyte Dataset* are iterated to determine the users with the highest and lowest five FFRs. In the event of a tie, the friend count will be used as the tie-breaking criterion. For the highest and lowest five users in the *10-Gigabyte Dataset*, their FFR, follower count, and friend count are tabulated. The user profiles of such users will be visited to qualitatively describe the users' characteristics.

RESULTS AND DISCUSSION: Table 4.14 highlights important metadata from the five users with the highest and lowest FFRs, as observed from the *10-Gigabyte Dataset*.

Looking at Table 4.14, among the users with a high FFR, `ConanOBrien`, `DalaiLama`, and `womensweardaily` are high-profile accounts (belonging to a celebrity, spiritual leader, and publication respectively). These three users have had their real-life identity verified by Twitter; such high-profile accounts have a '*verified*' logo on their Twitter web profile and have the metadata field `verified`= 1 (Section 4.3.2, also illustrated in Appendix A). These users fit the *broadcaster* stereotype cf. Krishnamurthy et al. [2008].

`Poconggg` is another popular, high-FFR-user fitting this stereotype, who is based in Indonesia and has gained a high number of Indonesian Twitter users who *follow* him; he did not, however, apply for verified status from Twitter. `SoalCINTA` on the other hand, is a novelty Twitter account which generates tweets on relationship advice in Indonesian, with a high FFR. Other users *follow* this account for its novelty tweets, fitting the `broadcaster` stereotype yet again[11].

[11] An interesting observation about `SoalCINTA`: when this user profile was accessed about four months after its initial discovery in the dataset, the number of followers have drastically dropped nine-fold to about over a hundred thousand. I suspect that there might either be a bug in Twitter's generation of the

Username	FFR	Follower count	Friend count	Notes
Top five FFRs				
ConanOBrien	4,330,976	4,330,976	1	Celebrity Conan O'Brien (*verified Twitter user*)
DalaiLama	3,099,603	3,099,603	0	Spiritual leader The Dalai Lama (*verified Twitter user*)
womensweardaily	1,892,270	1,892,270	0	Women's fashion website (*verified Twitter user*)
Poconggg	1,196,705	1,196,705	0	Indonesian writer Arief Muhammad
SoalCINTA	926,073	926,073	0	Indonesian Twitter account for relationship advice
Least five FFRs (sorted by decreasing friend count)				
khanjahed75	0	0	942	Suspicious account with links to unnamed videos.
sumikhatun37	0	0	935	(*Suspended account*)
YahooTflkkhan	0	0	932	(*Suspended account*)
saleha_soha	0	0	791	(*Suspended account*)
araitkh	0	0	636	Japanese user who frequently re-tweets messages/URLs.

Table 4.14: Statistics of users with the largest five FFRs (in descending order), and the lowest five FFRs (in decreasing order of friends).

At the other end of the spectrum, there were many users with the FFR of zero, i.e. the lowest FFR value possible. The ones mentioned in Table 4.14 were picked on the basis of a zero FFR, and having the highest number of friends. The top three accounts — sumikhatun37, YahooTflkkhan, and saleha_soha — were spam accounts that frequently posted spam links and exhibited aggressive *following* behavior [Thomas et al., 2011], as per the definitions of *miscreants* by Krishnamurthy et al. [2008]. Furthermore, when these user profiles on the Twitter website were accessed again, four months after their metadata was first read, I find that these accounts were removed by Twitter: similar to cases with spam users as seen in [Cheong and Lee, 2009, 2010b; Thomas et al., 2011]. khanjahed75 and araitkh also exhibit signs of *miscreant* or *evangelist* behavior [Krishnamurthy et al., 2008]; however these accounts are still accessible on Twitter at time of writing.

4.7.4 User Loyalty and Usage Frequency

Quantifying User Activity

User activity on a social network is a good way of characterizing or profiling a group of users. For example, Bozkir et al. [2010] have performed research on per-user daily

recorded metadata (when the metadata was first read), or that the followers who were removed en masse were actually comprised of spam accounts usage frequency and duration of each usage session with respect to Facebook demographic analysis (Section 3.5.1).

Examples specific to Twitter include the study of how crisis events can change one's messaging habits, e.g. in order to spread the word or to broadcast their current situation [Longueville et al., 2009; Herring et al., 2004]. Kwak et al. [2010] on the other hand performed a study on messaging frequency amongst a group of users with respect to a particular trending topic. In summary, these examples [Longueville et al., 2009; Herring et al., 2004; Kwak et al., 2010] focus on user activity on Twitter as a group with respect to a particular topic (crisis events/trends) found in the group's messages.

However, studies on individual Twitter user activity patterns, similar to Bozkir et al. [2010]'s work on Facebook users, are lacking in current research. Based on existing available user metadata on Twitter, I propose two new metrics designed to measure a user's activity and participation on Twitter.

Normalized Account Age

In the user domain, two metadata items are provided by the Twitter API: `statuses_count` (the total number of tweets the user has composed), and `created_at` (which stores the time-stamp the account was created). Using these pieces of metadata, one can obtain a user's *account age* [Cheong and Lee, 2010c]. *Account age* is defined as:

$$\text{Account age} = \text{Observation date} - \texttt{created\_at} + 1 \tag{4.3}$$

The account age is the number of days (in whole numbers) elapsed since a Twitter user's account has been created, up until a certain period, i.e. the *observation date.* The observation date refers to the day said user's metadata was accessed from the API: in other words, the date stamp which coincides with the time the tweet was broadcast via the Streaming API. The account age has been normalized by adding one to the date difference, to account for rounding up; e.g. an account six hours old is rounded up to be a day old.

The derived account age provides for the identification of new accounts and the degree of 'veterancy' of users [Cheong and Lee, 2010c]. Potential application for this include detecting user loyalty to Twitter as a medium of expressing oneself [Hughes and Palen, 2009], detecting opinion spam and sock-puppetry [Cheong and Lee, 2009; Barracuda Networks, Inc., 2010].

Messaging Frequency

A user's total number of statuses is also provided by the Twitter API as the metadata item `statuses_count`.

With this, I propose another metric, the *message frequency*, which is defined as [Cheong and Lee, 2010c]:

$$\text{Message frequency} = \frac{\Sigma\text{number of tweets posted } (\texttt{statuses\_count})}{\text{Account age}} \tag{4.4}$$

On Twitter, activity is characterized by the publication of tweets. Hence, I posit that messaging frequency reflects the degree of a user activity on Twitter in general [Cheong and Lee, 2010c]. In research, fluctuations in this figure can be used to look out for any undue influence that changes a users messaging rate over time, e.g. trending behavior or emergencies, cf. [Cheong and Lee, 2009, 2010c; Kwak et al., 2010; Huang et al., 2010; Kumar et al., 2010].

Case Studies on Messaging Frequency

As in the previous section, I will wrap up this section with a case study on how messaging frequency can be used to detect anomalies in Twitter user activity (Experiment 4.13.

Experiment 4.13. *To observe anomalies in the 10-Gigabyte Dataset by checking for users with anomalous messaging frequency scores.*

Method: The user records in the *10-Gigabyte Dataset* are iterated, and each one of their messaging frequencies calculated. The top five users with the highest messaging frequency are recorded, along with their total message count and normalized account age. Qualitative evaluation on these five users are performed by visiting their Twitter profiles and documenting any anomalies or peculiarities present.

Results and Discussion: Table 4.15 highlights statistics from five users with the highest messaging frequency as observed from the *10-Gigabyte Dataset*. The account age was measured by referring to the message datestamp during the period of data collection. A complete discourse and analysis of the overall distribution of user activity metrics in the dataset is included in the next chapter (Section 5.5).

From the statistics above, based on the messaging frequency alone, two classes of anomalous users can be identified:

1. **Spam users:** These users comprise of automated programs used to broadcast spam links en masse and sometimes exhibiting unusual *following* behavior (cf. Section 4.7.3). The abnormally high messaging frequency makes it unlikely that the tweets are composed by a user due to the time and effort involved. A proportion of these users are successfully flagged as spam accounts and consequently have their accounts suspended or banned by Twitter. Examples from Table 4.15 include `teamwfollowback`, `JenStar1`, `aurogWuB`, and `JooNelson2`.

2. **Novelty users:** These Twitter users, usually maintained by automated programs or groups, exist for the sake of novelty in that their tweets have a niche to them. An example in Table 4.15 is `ramedias`, purportedly based in Ukraine, which broadcasts up-to-the-minute temperature information as read from an input device.

4.8 Learning by Inference: Communication Patterns

I now shift my focus to studying metadata specifically found in the message domain, which opens up the study of messaging and communication patterns found on Twitter.

Username	Messaging frequency (messages per day)	Total messages	Norm. account age (days)	Notes
Top five users based on messaging frequency				
teamw-followback	7,254	21,762	1	(Suspended account) A novelty account which *follows* back any user who *follows* it, possibly used to inflate a user's number of followers.
JenStark1	2590.54	67,354	26	Suspicious account which tweets spam links, using automated Twitter spam-marketing software.
ramedias	2,218.19	155,273	70	Novelty account which tweets temperature readings from Ukraine, approximately once per minute
aurogWuB	2,209	2,209	1	(Suspended account) Spam account which tweets spam links padded with random text to make it seem legitimate.
JooNelson2	2172.72	63,009	29	Suspicious account which tweets messages repeatedly with hashtag #PodaSerMoraless, which aims to promote a Portuguese band[12].

Table 4.15: Statistics of users with the top five daily messaging frequencies (in descending order).

4.8.1 Message Length as an Indicator of Content

The length of a message is commonly used in investigating the socio-linguistic properties [Ling, 2005] of the users behind the tweets [Cheong and Lee, 2010c]. This identifies the difference in user information-sharing behavior; it can be used to differentiate e.g. users utilizing Twitter to broadcast long postings akin to conventional blogs, versus the broadcasting of short snippets to summon help or break news on the spot [Cheong and Lee, 2010c]. On Twitter, message length can be calculated by simply calculating the number of characters in the text metadata item, i.e. the actual message content itself. This is trivially done using string-length functions in most programming languages.

A fine example of related research is Yoshida et al. [2010] who looked at message length as one of the distinguishing factors in differentiating between bot-posted tweets and human-posted ones. Bot-generated tweets tend to have a higher average length and a distribution which is skewed towards the maximum 140-character limit, simply because they truncate long messages abruptly [Yoshida et al., 2010]. By comparison, humans will usually end their tweets as the length approaches the limit. Yoshida et al. [2010] also found in their studies that bimodal distribution of tweet lengths found in a 'natural' sample of tweets. A local maxima is initially observed in the distribution, which slowly tapers towards the 140-character limit, before suddenly spiking at the 140-character boundary. Such a bimodal distribution is characteristic of everyday tweets containing a dichotomy of both short and long messages. Such a distribution is also corroborated by surveys of the 'state of the Twittersphere', including Zarrella [2009]. The second spike is identified by Zarrella [2009] to consist of "...many users [reaching] the 140-character limit in an attempt to get as much content as possible into every update" [Zarrella, 2009].

4.8.2 Message Entities: Replies, Retweets, and Hashtags

A Primer on In-Text Entities

The introduction of *Entities* in Twitter message metadata and the Retweeting API (refer Section 4.3.1) allows easy access to three entities in the message text: `@user` replies, retweets, and hashtags.

To briefly recap, the definitions of the three entities (as per Section 4.3.1) are:

- *Reply messages:* messages preceded with `@user` to indicate a reply to another user.
- *Retweets* or *RTs*: messages of the format `RT @user message` to indicate forwarding of a message.
- *#Hashtags*: tweets with one or more keywords prefixed with a hash symbol (`#keyword`; the notation is used to 'tag' a tweet by keyword.

The introduction of such features to the Twitter API vindicates my proposal to single out the aforementioned three entities in initial studies; which I have documented in [Cheong and Lee, 2009] and [Cheong and Lee, 2010c].

Two Methods of Entity Extraction

However, as per my research [Cheong and Lee, 2009, 2010c,d, 2011] and those found in other prominent studies [Boyd et al., 2010; Honeycutt and Herring, 2009; Herring et al., 2004; Java et al., 2009], I focus instead on manual extraction of substrings from the `text` metadata. This is favored, as opposed to the new *Entities* and *Retweets* API, for two reasons:

1. **Textual representation**: Firstly, my approach merely requires the message text to be represented as an ASCII string, where string operations can be easily handled by most modern programming languages such as Perl or Java. This allows for adaptability of my approach to different programming languages and environments.
2. **Compatibility**: Secondly, this allows me to work with older datasets in 2009 when work on this PhD began — such as data extracted using the Twitter REST-`user` API — as such backdated data was created before the introduction of the new *Entities* and *Retweeting* features.

For the extraction of the three entities [Cheong and Lee, 2010c, 2011] from the message text, I will describe the character patterns specific to such entities in the form of a regular expressions, with a brief description of what each regular expression does:

- **Reply messages:** The regular expression `(\A|\s|\W)@(\w+)` detects the `@user` notation for any substring at the start of a message, or after whitespace, or after non-word characters.

- **Retweets:** The regular expression `(\A|\s|\W)RT(\A|\s)` detects the `RT` string for any substring at the start of a message, or after whitespace, or after non-word characters; AND followed by a whitespace or non-word character.

- **Hashtags:** The regular expression `((\A|\s|\W)#(\w+)` detects the `#hashtag` notation for any substring at the start of a message, or after whitespace, or after non-word characters.

Applications on Extracted Entities

Once these entities are extracted from the `message` text, I can then perform experiments on them, depending on the needs of a given study or experiment. Throughout this thesis, such approaches include:

1. **Determining the presence (or absence) of individual entities:** The presence or absence of particular entities in a message can be flagged by using a Boolean value. This is performed for case studies which apply pattern recognition algorithms on user and message inferences. Such studies are to be found in Chapters 5 and 6.

2. **Tabulating statistics on entities within a message:** This approach involves the construction of frequency distributions to map out the type of messages found within a given dataset. Again, this is trivially obtained: the extracted entities can simply be enumerated to obtain summary statistics such as number of entities per message, and average entities per message. This approach is used in Chapter 5.

3. **Isolating messages which contain specific entities:** By determining the presence of a particular entity, such as `#earthhour` (studied in Section 7.1), messages fitting a specific criterion can easily be singled out for further analysis.

Qualitative Studies on Entities

Analysis of these entities are well-documented in existing research, including my published studies [Cheong and Lee, 2009, 2010c, 2011, 2010b,d].

The following is a list of qualitative studies typically conducted on Twitter `message` entities:

- Patterns of interpersonal communication (via the `@user` directed reply notation) to users in a particular community especially during times of crisis [Honeycutt and Herring, 2009; Hughes and Palen, 2009; Huberman et al., 2008a].

- How a particular topic of interest rapidly gains popularity, using retweets [Cheong and Lee, 2009; Kwak et al., 2010].

- Information sharing and message dissemination via retweets [Boyd et al., 2010; Dearman et al., 2008].

- How tweets of a particular topic can evolve into conversational tags and micro-memes [Huang et al., 2010] by use of `#hashtags`.

- Searching for related tweets for extraction of additional information [Cheong and Lee, 2011; Boyd et al., 2010].

- Demographic analysis of the contributing user [Cheong and Lee, 2010c] using a combination of the three entities, as postulated in conventional blog research such as [Argamon et al., 2007; Herring et al., 2004].

4.8.3 Message Entities: URLs for Information Sharing

Information Sharing on Twitter via URLs

Besides the various entities discussed in the previous section, the tweet content itself can also contain links to other pages on the Internet. Prior to the introduction of the *Entities* and *Media* API extensions to the Twitter API (Section 4.3.1), there have been existing studies dealing with hyperlinks in message text. Examples of such studies include analysis of user intention via message content [Java et al., 2009], and analysis of information sharing patterns indicated by presence of URLs [Cheong and Lee, 2010c; Boyd et al., 2010].

One of the more comprehensive studies of links in tweet message content was performed by Yoshida et al. [2010] who analyzed the kind of URLs that are shared by users, and the effects of link sharing due to the differences in Twitter clients (see also Section 4.6.3).

Parsing URLs

The assertion of the presence of URLs in tweet messages can be done in either of two ways.

The first option is enumerating the new *Entities* metadata to check for mentions of `urls` to be extracted. The downside of this is that messages obtained prior to the introduction of *Entities*, or messages with discarded *Entities* metadata entries cannot be studied. Also, some URLs can be displayed on the main Twitter website in truncated form due to formatting issues, resulting a difference between the two types of URLs available in the `urls` entity.

Secondly, as espoused in existing studies, the `text` metadata item can simply be checked for the existence of the substring “`http://`”, indicating the start of a URL. Again, current programming languages such as Perl make this task trivial, by simply matching the regular expression `http://`.

Research on URLs in Tweets

From my research [Cheong and Lee, 2010c, 2011, 2010b], the presence of URLs indicates an intention by the user to share information, not unlike behavior exhibited on *de facto* social networks and news aggregators. A URL provided in a tweet usually complements the current discussion within a tweet with other news sources and media coverage [Cheong and Lee, 2011], or simply to 'convey larger amounts of information' such as in forum posts or conventional blog posts due to the constraint of the 140-character limitation [Hughes and Palen, 2009; Starbird et al., 2010].

Specifically, one is also able to study the presence of media sharing — commonly photos and videos — via Twitter. Sites such as *TwitPic*, *Imgur*, and *Flickr* can be used to share pictures with other Twitter users. Videos on the other hand can be shared using sites such as YouTube.

By analyzing the URL substrings contained within messages on a given topic, the presence of auxiliary user-generated content can be determined. Potential applications of the study of URLs in tweets include:

- Studying how auxiliary media is shared using Twitter as the medium [Dearman et al., 2008; Schrammel et al., 2008].

- Obtaining first-hand information (e.g. eyewitness videos and pictures) during disaster and crisis situations [Hughes and Palen, 2009; Fleishman, 2009; Terdiman, 2009].

- Chronicling crisis or terrorism events based on civilian reaction [Cheong and Lee, 2011], as seen in recent terror events [Beaumont, 2008; Cashmore, 2009a]; sources such as these are a wealth of information to authorities seeking to chronicle such activities [Cheong and Lee, 2011].

4.8.4 Applications of Entity Analysis, Visualization and Graphing

As the essence of a tweet's content — i.e. the text message metadata item studied in this section — is simply an unstructured text string, existing methods of textual analysis can directly be applied. As documented in Section 3.4, research studies on modeling [Ritter et al., 2010], personalized user recommendations [Bernstein et al., 2010; Phelan et al., 2009; Wu et al., 2010], sentiment detection [Lee et al., 2010; Wasow and Baron, 2010; Shamma et al., 2009, 2010] and user search [Suh et al., 2010; Golovchinsky and Efron, 2010], among others, have been successfully applied to tweet content.

However, as the scope of this chapter (and this thesis in general) is about the discovery of inferences from metadata, I would not be delving into such topics. Rather, I will demonstrate some approaches to summarizing and visualizing tweet content, borrowing from the field of information retrieval, with special emphasis given to *entities* as discovered in recent sections. I would like to emphasize that such information retrieval-based approaches are not new; rather, the adaptation of such approaches on tweet entities and text constitute the discussion in this section.

Keywords: Vocabulary Analysis

Before any lexical analysis of a tweet can take place, it needs to be separated into its constituent words. Splitting the words from tweets can be trivially implemented using a *string tokenizer* found in most programming languages. The tokens are then normalized based on case (e.g. normalizing everything to lowercase), and sanitized to remove non-alphanumeric characters (e.g. punctuation, HTML entities, and special characters).

The resulting list of words will still contain a variety of *stopwords* [Luhn, 1958]: a set of commonly-occurring function words such as articles (*an*, *the*), pronouns (*her*, *me*), particles (*if*, *however*), and conjunctions (*and*, *but*). These stopwords do not contribute to the lexical content of the text — i.e. "...[lexical] significance sought here does not reside in such words" [Luhn, 1958] — but rather skew the frequency distribution of unique words due to their ubiquitous presence [Russell, 2011b]. Removing these stopwords from the original list of words would yield a subset of words which are more significant [Luhn, 1958] in capturing the essence of the tweet.

Luhn's Summarization and Frequency-based Analysis

The usage of Luhn summarization on social media including tweets have already been documented in e.g. [Russell, 2011a,b], and is included here for the sake of brevity. Luhn's idea of summarization, based on the "frequency analysis of the words in [a] document" [Russell, 2011a], has been proposed in his seminal paper [Luhn, 1958]. In the context of tweets, the "document" refers to a collection of one or more tweets to be studied. Frequency-based analysis simply consists of "[taking] an *inventory...* and [having] a word list compiled *in descending order of frequency*" [Luhn, 1958] (emphases mine).

In the case of Twitter, the only major difference between conventional Luhn summarization is the inclusion of entities, whereby `@user`, `RT`, and `#hashtags` are to be considered verbatim as unique keywords in the frequency inventory [Russell, 2011b]. URLs are usually excluded, though extra processing to extract only the domain names is possible (cf. algorithms in Section 4.7.2). With the generated inventory of unique keywords (including entities) and their frequencies, it is then trivial to obtain a user's *lexical diversity* for a particular collection of tweets: i.e. the ratio of total unique keywords over the total word count [Wagner and Strohmaier, 2010; Russell, 2011b]. Simply put, lexical diversity is a simple measure of the size of the vocabulary for the author(s) of a given set of tweets.

Among examples of research in the applications of Luhn frequency-based tweet analyses are:

- Determining the most commonly occurring Twitter user mentions (`@user` entities) and themes mentioned within tweets (`#hashtags`) in the 2011 London Riots (Section 7.3, published as [Cheong et al., 2012a]).

- Determining the most likely areas in Australia (using location `#hashtags`) where Twitter participation on the yearly Earth Hour event is most frequent (Section 7.1, published as [Cheong and Lee, 2010d]).

- In marketing, analyzing the keywords and frequent `#hashtags` present in tweets for different beer companies, to determine common marketing themes [Watne and Cheong, 2012]. Lexical diversity was also measured for each company's Twitter account, given a Luhn frequency table [Watne and Cheong, 2012].

Visualization of Luhn Frequency Tables using Tag Clouds

With the creation of a Luhn frequency table, one can also easily visualize the list of unique keywords (and/or entities) based on their frequency within tweets. A *tag cloud* refers to the arrangement of words, where the importance of each word is distinguished by e.g. font size, weight, and color [Halvey and Keane, 2007].

As the use of tag clouds to visualize tweet keywords has already been discussed in Chapter 3, evaluated in current literature e.g. [Russell, 2011a; Bernstein et al., 2010; Donath et al., 2010; Bloch and Carter, 2009; Mendoza et al., 2010], and beyond the scope of this section, I will not include an in-depth coverage of tag cloud construction and technical details on its implementation.

Co-occurrence Networks for Graph-theoretic Analysis of Keywords/Entities

Before I conclude this chapter, I would like to draw attention to the concept of *co-occurrence networks based on tweet keywords/entities*, another promising area of keyword and entity analysis for tweets (in the message domain). Drawing from the research proposed by Z. Fan and D. S. Stones [pers. comm., 26 September 2012], it is possible to construct a "co-occurrence network" for a particular keyword (or entity) found in a set of strings, by constructing a network from which the nodes are made up of other keywords (entities) that occur within a same tweet.

Although still a work in progress, and hence beyond the scope of this thesis, the research methodology by Fan & Stones [pers. comm., 26 September 2012] can directly be applied on tweets. This line of research is useful for spatial visualization of how keywords/entities related to a particular topic/user are distributed; and the *network fingerprinting* of tweet keywords/entities from a graph-theoretic perspective.

4.9 Concluding Notes

From this chapter, I have first detailed the inner workings of the Twitter API for collection of metadata from Twitter. In it, I have detailed the strengths and weaknesses of the on-demand APIs (REST-`user` and `search`) compared to the Streaming API.

The notion of duality of the original Twitter APIs has led me to propose that both the Twitter domains — of users and messages — need to be studied in conjunction with one another. By revisiting the literature discussed throughout Chapter 3, I have identified that research on identifying latent trends and inferences from metadata from both domains are lacking in the current state-of-the-art.

After briefly discussing the useful pieces of metadata extractable from Twitter's APIs, I presented my research contribution in the form of ten new inference algorithms or metrics. These algorithms work on both user and message metadata to uncover real-life demographic properties among Twitter users, features of a Twitter user's online presence, and also tweeting habits and eccentricities.

I provided evaluations of each algorithm's performance on the real-world *10-Gigabyte Dataset*, which is of the order of millions of records. Discussion on the observed results, practical applications on real-world data, and case studies were provided alongside with each algorithm/metric.

The following chapter will follow naturally from the results of this chapter. Chapter 5 will initially discuss two prototypes of my design: a large-scale data collection framework via the Streaming API, and a smaller-scale data collection framework [Cheong and Lee, 2010c] that have worked on the on-demand APIs (REST-`user` and `search`) before they were superseded by the Streaming API. The latter part of Chapter 5 will focus on the applications of my algorithms/metrics (introduced in this chapter) on the entirety of the *10-Gigabyte Dataset*; followed by a discussion on the real-world properties of Twitter users as seen in the end results of the algorithms/metrics.

Chapter 5

Analyzing Large-Scale, Real-World Twitter Data

> "*Woke up this morning, from the strangest dream,*
> *I was in the biggest army, the world has ever seen...*"
>
> — Hunters and Collectors,
> *Holy Grail* (1992)

Parts of this chapter have been published as:

Cheong, M. and Lee, V. [2009]. Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base, Proc. CIKM 2009 Co-Located Work- shops: SWSM 2009, pp. 1–8.

Cheong, M. and Lee, V. [2010c]. Twitmographics: Learning the Emergent Properties of the Twitter Community, *From Sociology to Computing in Social Networks: Theory, Foundations and Applications, Vol. 1 of Lecture Notes in Social Networks*, Springer-Verlag, pp. 323–342.

Cheong, M. and Lee, V. [2011]. A Microblogging-based Approach to Terrorism Informatics: Exploration and Chronicling Civilian Sentiment and Response to Terrorism Events via Twitter, Information Systems Frontiers **13**(1): 45–59.

Cheong, M., Ray, S. and Green, D. [2012a]. Interpreting the 2011 London Riots from Twitter Metadata, *Proc. SoCPAR 2012.*

In the previous chapter, I illustrated that Twitter is a rich source of data for research. In particular I described the types of metadata available on Twitter from both the user and message domains; and described how one can make sense of such data to reveal interesting properties. This chapter naturally follows from the previous one by elaborating how my contributions from the previous chapter can be applied in the real-world.

Firstly in Sections 5.1 and 5.2, I will demonstrate several prototypes for metadata harvesting frameworks, for both the superseded Twitter REST (On-demand) API, and the newer Twitter Streaming API. Section 5.1 documents a framework [Cheong and Lee, 2010c] for the on-demand API, developed and published in the earlier stages of my research. This section is followed by my next contribution: a Streaming API-based metadata harvester (Section 5.2) for data collection on a grand scale. This latter prototype was used successfully in a large-scale data gathering exercise, resulting in the *10-Gigabyte Dataset* alluded to in Section 4.5, which will be documented fully in 5.3.

The latter sections (Sections 5.4, 5.5 and 5.6) highlight my contributions to the large-scale analysis of, and observations on the *Twitterverse* circa 2011–2012. This is achieved by applying my novel inference algorithms and metrics, introduced in Chapter 4, on the *10-Gigabyte Dataset* gathered using the framework in Section 5.2. Section 5.4 outlines the real-world demography of Twitter users from the *10-Gigabyte Dataset* in terms of gender, geographic location, and device classes used. Section 5.5 reveals the online presence properties of Twitter users, thanks to my algorithms in Section 4.7. The last section of this chapter, Section 5.6, deals with an analysis of communication/tweeting patterns in Twitter in terms of the summary statistics and entities used in all tweets in the *10-Gigabyte Dataset*. The novelty of my contributions in this chapter lies in the fact that such applications of inference and knowledge algorithms on Twitter data of such width and depth is few and far between in extant research.

5.1 *Twitmographics*: On-Demand Metadata Harvesting

In the early stages of my PhD research circa 2009, while investigating methods of metadata extraction from Twitter, I have designed and published a framework for automated harvesting of data from Twitter using the Twitter on-demand APIs (REST-`user` and `search`) [Cheong and Lee, 2010c]. As this framework — *Twitmographics* [Cheong and Lee, 2010c] — utilized the old REST-`user` and `search` APIs, there were inherent limitations such as:

- **Inconsistency of retrieved data**: Due to differing API limits, the number of messages versus the number of users can differ (Section 4.1.3);
- **Continuity and volume of data**: The volume of messages can be inconsistent due to the design limitations of the `search` API; and
- **Lack of new functionality**: Enhanced location metadata (Section 4.6.2), entities (Section 4.8.2), and other new metadata fields were not available from Twitter during development of this prototype.

Nonetheless, this section introduces the design of my *Twitmographics* framework in terms of overall structure, connectivity to Twitter, the types of metadata fields extracted and analyzed, and some sample output generated by this framework.

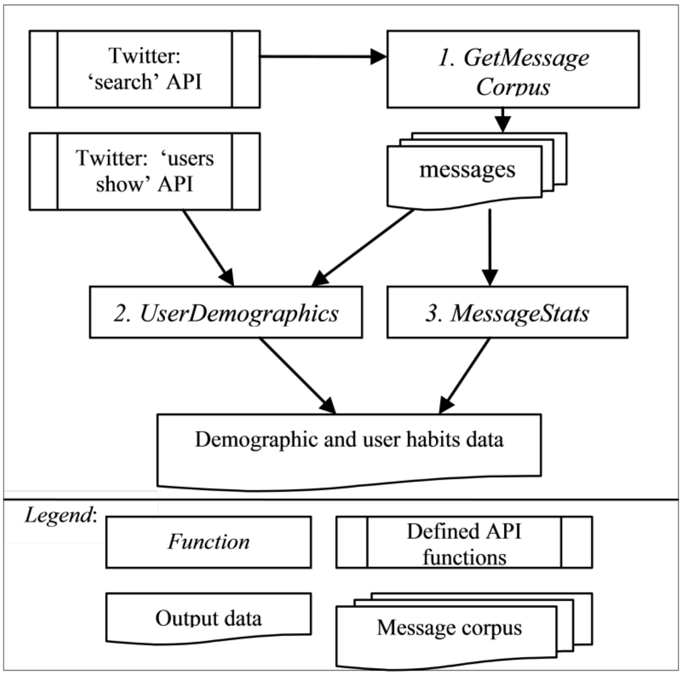


Figure 5.1: The basic structure of the *Twitmographics* framework, as published in 2009 [Cheong and Lee, 2010c].

5.1.1 Design Overview

Perl was used in the framework's overall development as it is well-suited to processing large chunks of records and textual information. The `Net::Twitter` module[1], a wrapper that provides the necessary low-level functionality for communication to Twitter's servers and accessing of the Twitter APIs.

The overall framework of *Twitmographics* is presented in Figure 5.1, wherein the following three distinct processing modules are labeled in italics:

- **GetMessageCorpus** searches for the relevant messages based on a specified search query, and saves them on disk for further processing.
- **MessageStats** reveals message statistics, embedded in the metadata of a Twitter message.
- **UserDemographics** provides the underlying emergent properties from the user base, i.e. the authors of the messages harvested in *GetMessageCorpus*.

5.1.2 *GetMessageCorpus* module

The first processing module in *Twitmographics* — *GetMessageCorpus* — harvests raw message data by querying the `search` API for mentions of a topic.

The messages returned by the `search` API will be cached in memory, and also saved to disk for further analysis. There is a potential for messages to be duplicated across queries to the `search` API; hence this module will discard any results that have been hitherto

[1] CPAN page for Marc Mims' `Net::Twitter`: <`http://search.cpan.org/~mmims/Net-Twitter/`>

analyzed during a previous run. The raw data obtained from this module is of little value for processing; which is where the *MessageStats* and *UserDemographics* modules come into play. The pseudocode in Algorithm 5.4 briefly describes the functionality behind this module.

Algorithm 5.4 The *GetMessageCorpus* module in the *Twitmographics* prototype.

```
procedure GetMessageCorpus(topic)
    results ← query for topic in SearchAPI
    for all message in results do
        messagecorpus ← messagecorpus + message
            ▷ append message to the messagecorpus in memory (as a hash) and on disk (as a flat file dump)
    end for
end procedure
```

5.1.3 *MessageStats* module

The second major component in *Twitmographics* is *MessageStats*, which performs preliminary analysis of the raw message metadata obtained by *GetMessageCorpus*. The ideas behind this module are illustrated in Algorithm 5.5.

Algorithm 5.5 The *MessageStats* module in the *Twitmographics* prototype.

```
procedure MessageStats(messagecorpus)
    for all message in messagecorpus do
        device_platform ← match message source with software name
        content_length ← parse length of message text
        content_features ← perform regular expression matching on message text
        UserDemographics(message)
    end for
end procedure
```

As this prototype was developed in the early days of my PhD research [Cheong and Lee, 2010c], the *MessageStats* module in *Twitmographics* returns only seven attributes. These are merely a subset of the entire list of inferences as defined in Sections 4.6, 4.7, and 4.8. Table 5.1 details the seven inferences, as they exist in the *Twitmographics* prototype and the algorithmic processing involved.

5.1.4 *UserDemographics* module

The third major component in *Twitmographics* is *UserDemographics*, which further obtains user information based on a tweet's username. The username is extracted from the message (via *MessageStats*); this is followed by the downloading of user information via the Twitter REST-`user` API for further analysis. Algorithm 5.6 illustrates the concepts behind *UserDemographics*.

Again, as *Twitmographics* was devised in the initial stages of research, the inferences and attributes generated merely form a subset of Sections 4.6, 5.5, and 5.6. Table 5.2 details the eight attributes, as they exist in the *Twitmographics* prototype.

Table 5.1: The seven *message inferences*, as used in the development of *MessageStats* in the *Twitmographics* prototype.

Inference	Description
Device/platform classification:	Twitter messages are classified as belonging to one of seven distinct classes of devices and software clients: 1. the official Twitter web interface 2. mobile devices 3. social media applications 4. alternative Twitter software clients 5. feed aggregators 6. Twitter mash-ups for information sharing and 7. Twitter marketing and bulk messaging tools In *Twitmographics*, the pool of *client IDs* contained 66 unique pieces of software. The device and platform categories are ascertained by the software authors' descriptions in their websites, in my first attempt [Cheong and Lee, 2010c] to extend the categorization performed in prior work [Krishnamurthy et al., 2008]. Initial applications of device/platform classification include the detection of censorship; and determining the reach of e.g. mobile and ubiquitous computing versus computer-based Twitter usage.
Message length (Section 4.8.3)	In this prototype, I discretized the message length into one of 14 separate bins, each bin in multiples of ten characters.
Presence of `@user` replies (Section 4.8.1)	This is a Boolean value indicating a reply-based `@user` tweet.
Presence of retweets / `RTs` (Section 4.8.1)	This is a Boolean value, indicating the presence of the retweeting (`RT`) token.
Hashtagging behavior (Section 4.8.1)	This is a Boolean value, indicating the presence of the `#hashtag` notation, used to socially tag a tweet.
Presence of URLs (Section 4.8.3)	This is a Boolean value that shows if URLs are present, alluding to sharing of information, similar to what happens in social networks and news aggregators.
Picture attachments	In my preliminary study of further message inferences [Cheong and Lee, 2010c], the presence of the term `twitpic` — a photo-sharing service [Twitpic Inc., 2009] — is an indicator of the presence of linked images in Twitter message contents. This is synonymous with user-generated content sharing and reporting of eyewitness news. Again, one could allude to the presence of these linked images an indicator of a particular user (or community's) need for computer-mediated information sharing [Dearman et al., 2008; Hughes and Palen, 2009; Schrammel et al., 2008].

Table 5.2: The eight *user inferences*, as used in the development of *UserDemographics* in the *Twitmographics* prototype.

Inference	**Description**
Gender (Section 4.6.1)	Gender is one of the attributes that can hint on a user's identity. To identify the gender of a Twitter user, I used my probabilistic ranking algorithm (Algorithm 4.1) [Cheong and Lee, 2009] to process the users' given name using the 1990 Census rank data of 5,494 unique first names (Section 4.6.1).
Location (Section 4.6.2)	At time of *Twitmographics*' development, Twitter has only two pieces of data to identify a users' location, both of which are located in the user `location` field. This is due to unavailability of the Places and Geo API (cf. Section 4.1.4) before late 2009. This prototype supports the processing of either free-form location strings, or specially-formatted coordinates by GPS-enabled mobile clients. The Google Maps Geocoding API was used due to the small size of data available, i.e. the order of thousands of records (Section 4.6.2).
Web usage habit/generalization (Section 4.6.3)	The Twitter profile page for a user also lets the user publish his/her website's URL. *Twitmographics* classified a user's profile URL into one of nine *usage stereotypes*, defined prior in Table 4.12 (Section 4.8.1).
Profile picture (avatar) presence (Section 4.7.1)	In the *Twitmographics* prototype, I check only for the existence (or absence) of a user's profile picture using contents of the `profile_image_url` field. Twitter users who choose to put profile/avatar pictures are more likely to interact and participate in Twitter activity as opposed to those who do not.
Follower/friend ratio, FFR (Section 4.7.3)	This simple metric was used in *Twitmographics* as an indicator of the dynamics of networking and interaction for a particular user. The FFR is simply the quotient of $\frac{\texttt{followers\_count}}{\texttt{friends\_count}}$ (Equation 4.2). I discretized this ratio into one of the following seven groups, similar to the strategy used by Barracuda Networks, Inc. [2010]: 1. Famous user/information source: FFR > 4.00 2. Twice as many followers: $2.00 \leq$ FFR ≤ 4.00 3. Slightly more followers: $1.10 \leq$ FFR ≤ 2.00 4. Balanced: $0.90 \leq$ FFR ≤ 1.10 5. Slightly more friends: $0.50 \leq$ FFR ≤ 0.90 6. Twice as many friends: $0.25 \leq$ FFR ≤ 0.50 7. Unpopular user/information sink: FFR < 0.25

Inference	**Description**
Account age (Section 4.7.4)	*Twitmographics* also factors in a user's account age, characterized by the difference (in days) between the user's latest tweet and the day the user account was created (Equation 4.3). This allows identification of new accounts, checking the degree of 'veterancy' of users, estimating usage frequency and 'user loyalty' for Twitter, and detecting opinion spam or 'sock-puppetry' [Pang and Lee, 2008], as found in my earlier paper [Cheong and Lee, 2009]. The account age is quantified into one of seven possible groups: 1. Fresh user: less than a day; 2. Within a week; 3. Within a month; 4. Within a quarter (3 months); 5. Within a half-year (6 months); 6. Within a year; and 7. More than a year.
Message frequency (Section 4.7.4)	Messaging frequency for a given user is simply defined as per Equation 4.4: $\text{Message frequency} = \dfrac{\Sigma\text{number of messages posted}}{\text{account age (in days)}}$ The message frequency allows one to see a user's typical frequency of Twitter participation and detect any undue influence a particular topic might have in increasing this frequency (for example trending behavior, memes, or emergencies). *Twitmographics* discretizes the frequency into one of several groups: 1. less than 2; 2. 2 or more but less than 5; 3. more than 5 but less than 10; 4. more than 10 but less than 25; 5. more than 25 but less than 50; and 6. 50 or more.

Inference	**Description**
Violation of terms of service.	Another piece of information vital to understanding the Twitter user base is the presence of banned or deactivated accounts due to violation of terms of use, which was first applied in Cheong and Lee [2009]. The presence of such accounts, despite their scarcity, are meaningful as I could accurately pinpoint users who might have 'polluted' the Twitter message stream with spam, misleading messages, and (in several cases) scamming and phishing tweets [Thomas et al., 2011; Cheong and Lee, 2010c].

Algorithm 5.6 The *UserDemographics* module in the *Twitmographics* prototype.

```
procedure UserDemographics(message)
    username ← extract username field from message
    if username has been cached in hashtable then
        demographics ← look up username in hash
        return demographics
    end if
    metadata ← query for username via UserAPI
    if username is banned according to UserAPI then
        demographics ← {}
        return with error due to banned user account
    end if
    gender ← run ranking algorithm on metadata
    country ← run Google Geocoder API on metadata
    web_usage_habits ← perform string processing on metadata
    Twitter_usage_habits ← perform statistical calculations on metadata
    demographics ← {gender,country,web_usage_habits,Twitter_usage_habits}
    Cache demographics in hashtable and save to disk
    return demographics
end procedure
```

5.1.5 Message Harvesting Process

In *Twitmographics*, the Twitter on-demand APIs — specifically `search` API and REST-`user` API — are used. As *Twitmographics* was developed in mid-2009 [Cheong and Lee, 2010c] before Twitter Inc. strictly enforced rate-limits (Section 4.1.3), the amount of metadata records that could be retrieved were only in the order of thousands.

For the purposes of this prototype, I requested white-listing permission from Twitter Inc. which was still available during *Twitmographics*' creation. This allowed a maximum of 20,000 pieces of unique user information per hour, via the REST-`user` API. However, the 1,500-message search limit (Section 4.1.3) was imposed on the `search` API.

The workaround was to perform several search operations spaced in an interval of approximately 10 minutes between each run. This enabled *Twitmographics* to retrieve tens of thousands of messages per hour, corresponding with the maximum user data retrieval limit of 20,000 per hour as stated [Cheong and Lee, 2010c].

5.1.6 Sample ***Twitmographics*** Output

This prototype stores its output in plain text files, with records separated with newlines (`\n`), and fields within a record separated with commas (`,`). The output comes in two forms:

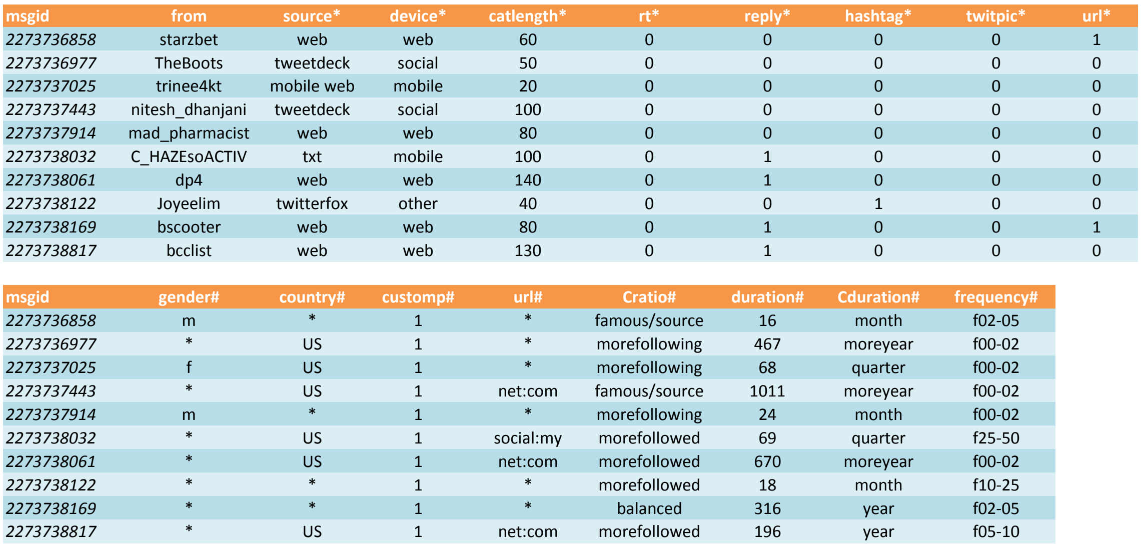
1. **raw metadata** as obtained from the REST-`user` and `search` APIs (e.g. `from`, `source`).

2. **inferences/attributes** in Tables 5.1 and 5.2, obtained after post-processing by *MessageStats* and *UserDemographics* respectively (e.g. *gender*, *country*).

Figure 5.2 illustrates sample output from the prototype. The input query `iPhone`, which coincided with the launch of the iPhone 3.0 software launch, was provided as input to *Twitmographics* [Cheong and Lee, 2010c]. The sample output in Figure 5.2 is based on experiments conducted in Section 6.3; published as [Cheong and Lee, 2010b] and [Cheong and Lee, 2010c]. The columns represent the attributes and raw metadata obtained; each row represents a record for each message (and its author). The table is split into two parts for legibility; the second table is merely a continuation of the first (rows indexed by the message ID in the first column).

I will not detail the complete analysis of the output data here, as the `iPhone` case study will be the focus of discussion in Section 6.3. For clarity, however, I will explain features of the sample output from the point of view of an individual record. Take the first row for example, with the message ID (`msgid`) of `2273736858`:

- The tweet was authored by user `starzbet` using the `web` Twitter client (`source` string), which is classified as the `web` *device class*. The next few columns indicate that the message is in the 60 character length group (60-69, inclusive); is neither a retweet, nor a reply message, nor contains a hashtag (`rt*=0`, `reply*=0`, `hashtag*=0`). However, the message does contain information-sharing properties due to inclusion of an URL (`url*=0`).



msgid	from	source*	device*	catlength*	rt*	reply*	hashtag*	twitpic*	url*
2273736858	starzbet	web	web	60	0	0	0	0	1
2273736977	TheBoots	tweetdeck	social	50	0	0	0	0	0
2273737025	trinee4kt	mobile web	mobile	20	0	0	0	0	0
2273737443	nitesh_dhanjani	tweetdeck	social	100	0	0	0	0	0
2273737914	mad_pharmacist	web	web	80	0	0	0	0	0
2273738032	C_HAZEsoACTIV	txt	mobile	100	0	1	0	0	0
2273738061	dp4	web	web	140	0	1	0	0	0
2273738122	Joyeelim	twitterfox	other	40	0	0	1	0	0
2273738169	bscooter	web	web	80	0	1	0	0	1
2273738817	bcclist	web	web	130	0	1	0	0	0

msgid	gender#	country#	customp#	url#	Cratio#	duration#	Cduration#	frequency#
2273736858	m	*	1	*	famous/source	16	month	f02-05
2273736977	*	US	1	*	morefollowing	467	moreyear	f00-02
2273737025	f	US	1	*	morefollowing	68	quarter	f00-02
2273737443	*	US	1	net:com	famous/source	1011	moreyear	f00-02
2273737914	m	*	1	*	morefollowing	24	month	f00-02
2273738032	*	US	1	social:my	morefollowed	69	quarter	f25-50
2273738061	*	US	1	net:com	morefollowed	670	moreyear	f00-02
2273738122	*	*	1	*	morefollowed	18	month	f10-25
2273738169	*	*	1	*	balanced	316	year	f02-05
2273738817	*	US	1	net:com	morefollowed	196	year	f05-10

Figure 5.2: Sample output of the *Twitmographics* prototype, given the search query `iPhone`.

- The user is a male (`gender#=m`), has customized his Twitter profile (`customp#=1`), and has an FFR greater than 4.0 (`Cratio#=famous/source`). As for his Twitter usage behavior, his account was 16 days old during the observation period (his account age will be in the `Cduration#=month` category), and his daily frequency of tweeting is (`frequency=02-05`). However, due to the lack of information, I could neither deduce the user's location, nor his web presence based on profile URL.

5.2 *TweetHarvester*: Real-time Metadata Acquisition

The latter part of my PhD research in 2010–2012 focused on the usage of the Streaming API (for reasons mentioned in Section 4.1.3). The Streaming API, as opposed to the hitherto-used REST-`user` API, necessitates a different approach to tweet harvesting and post-processing, highlighted in Section 4.2.1. An example is the difference in magnitude of data that can be obtained, which requires post-processing to be done separately from data harvesting.

Hence, I developed another framework similar to *Twitmographics* to harvest data from the Streaming API instead. This prototype system, nicknamed *TweetHarvester* as it harvests streaming tweets in real-time, consists of several Perl programs:

- ***HarvestSpritzer***: Establishes a socket connection to the Twitter Streaming API's `sample` method (which Twitter codenamed `spritzer`). After the connection is established, it continuously listens to a sampling of all public tweets, gathers metadata on the message and corresponding user, and dumps them to disk.

- ***HarvestFilter***: Similar to `HarvestSpritzer`, this Perl script also connects to the Streaming API via sockets, but uses the API's `filter` method instead. This allows one to narrow the incoming tweet stream based on a given search criteria instead.

- ***RawStreamer***: Using similar principles to *HarvestSpritzer*, *RawStreamer* is a real-time visualization tool for incoming Twitter messages from the `spritzer` method.

5.2.1 Prototype Design

This collection of Perl scripts originally started as *RawStreamer*, as an experiment in late 2009 to test the (then fledgling) Streaming API. As I gradually focused on the Streaming API's suitability for research, this script was further developed to allow storing of incoming tweets and metadata to disk. However, this turned out to be computationally expensive, due to the sheer volume of data that need to be processed.

During initial experimentation, a throughput of a few hundred messages per minute was achieved using the Streaming API, with white-listing permissions from Twitter Inc. However, as Twitter began changing the policy of white-listing and enforcing rate-limiting on the on-demand APIs (see also Section 4.1.3), the throughput of tweets through the Streaming API has changed from its earlier limit. As of the time of writing, 1% of 140 million tweets daily [Twitter Inc., 2012a] amounts to a throughput of approximately ∼1,000 tweets per minute.

RawStreamer, an old prototype created from the ground up based on raw socket programming –as it currently stands — is used purely as a visualization tool built using SDL, a OpenGL-like graphics library for Perl. Incoming tweets from the Streaming API are printed on-screen, with summary statistics from tweet metadata displayed in a separate column. For historic purposes, I have provided a brief overview of the design and usage of *Streamer* in Appendix C.

To allow robust and speedy harvesting of the Streaming API based on the new volume of data, the content harvesting and data dumping functionality was migrated from *Streamer* into the standalone *HarvestSpritzer* program. The extracted socket connection and raw data reading code — originally written from scratch as part of the initial *Streamer* design — has been removed due to performance issues. The newly refactored program was rebuilt using `AnyEvent::Twitter::Stream`, an optimized and stable Perl wrapper for the Twitter API[2].

HarvestFilter has the same code base as *HarvestSpritzer*, modified to use the `filter` method instead. All these scripts in the *TweetHarvester* prototype depend on an auxiliary library — `HUtils` — that incorporates all the knowledge discovery and inference algorithms described in Sections 4.6 through 4.8, and also functions that store associated metadata on disk. `HUtils` also contains various helper methods such as those involved in data sanitization, simple calculations, and file input/output.

5.2.2 Algorithmic Overview

This section details the inner workings of the *HarvestSpritzer* (and related *HarvestFilter*) prototypes, written in Perl. Algorithm 5.7 details the logic behind *HarvestSpritzer*.

[2] CPAN page for Tatsuhiko Miyagawas `AnyEvent::Twitter::Stream`: <`http://search.cpan.org/~miyagawa/AnyEvent-Twitter-Stream`>

Algorithm 5.7 The *HarvestSpritzer* algorithm.

```
procedure HarvestSpritzer
    outputdirectory ← directory on disk to store harvested raw data
    starttime ← current time
    interval ← interval before flushing buffer to disk
    limit ← maximum number of records to be retrieved
    filehandle ← open write handle for text file named starttime in outputdirectory
    write header row to filehandle
    streamobject ← new AnyEvent::Twitter::Stream reader object
    initialize streamobject with Twitter login credentials
    initialize streamobject with Streaming API mode spritzer
            ▷ this causes the API to return a sample of all tweets, as per Section 4.2.1
    establish streamobject connection to Streaming API
    begin busy wait on streamobject for tweets
    for all tweetrecord obtained from streamobject do
        counter ← counter + 1
        if tweetrecord is blank then
            next
        end if
        buffer ← buffer + sanitized, tab-separated message metadata
        buffer ← buffer + sanitized, tab-separated user metadata
        buffer ← buffer + sanitized entities metadata
                ▷ the new entities metadata items in Twitter, as per Section 4.3.1.
        buffer ← buffer + record separator (\n)
        if (counter is a multiple of interval) or (counter = limit) then
            flush buffer to filehandle
            clear buffer
        end if
        if counter = limit then
            close filehandle
            return
        end if
    end for
end procedure
```

For brevity, I have omitted the technical information behind the low-level socket programming used (used in establishing connections to the Twitter servers and authentication). It is available for reference in Appendix C.

As for the *HarvestFilter* prototype, its inner workings are generally similar to *Harvest-Spritzer*, save for one key difference — a different Streaming API method, `filter` is used instead of `spritzer`. Algorithm 5.8 illustrates this difference, where the `filter` method and the `track` parameters are used in *HarvestFilter* instead (Algorithm 5.8: lines 10–11 inclusive).

Algorithm 5.8 The *HarvestFilter* algorithm.

```
 1: procedure HarvestFilter
 2:     $outputdirectory \leftarrow$ directory on disk to store harvested raw data
 3:     $starttime \leftarrow$ current time
 4:     $interval \leftarrow$ interval before flushing buffer to disk
 5:     $limit \leftarrow$ maximum number of records to be retrieved
 6:     $filehandle \leftarrow$ open write handle for text file named starttime in outputdirectory
 7:     write header row to filehandle
 8:     $streamobject \leftarrow$ new AnyEvent::Twitter::Stream reader object
 9:     initialize streamobject with Twitter login credentials
10:     initialize streamobject with Streaming API mode filter
11:     initialize streamobject with search query as comma-separated track parameters.
12:         ▷ this causes the API to return only tweets containing track-ed keywords, e.g.
    “coffee,tea”
13:     establish streamobject connection to Streaming API
14:     begin busy wait on streamobject for tweets
15:     for all tweetrecord obtained from streamobject do
16:         $counter \leftarrow counter + 1$
17:         if tweetrecord is blank then
18:             next
19:         end if
20:         $buffer \leftarrow buffer +$ sanitized, tab-separated message metadata
21:         $buffer \leftarrow buffer +$ sanitized, tab-separated user metadata
22:         $buffer \leftarrow buffer +$ sanitized entities metadata
23:                   ▷ the new entities metadata items in Twitter, as per Section 4.3.1.
24:         $buffer \leftarrow buffer +$ record separator (\n)
25:         if (counter is a multiple of interval) or ($counter = limit$) then
26:             flush buffer to filehandle
27:             clear buffer
28:         end if
29:         if $counter = limit$ then
30:             close filehandle
31:             return
32:         end if
33:     end for
34: end procedure
```

5.2.3 Output Format

As described earlier in Algorithms 5.7 (line 2), and Algorithm 5.8 (line 2), the prototypes above store the collected data in a specified output directory.

To simplify post-processing, data is stored in ASCII-encoded flat text files; where non-ASCII characters are sanitized before storage. I use the newline character (`\n`) as the record separator, and the tab character (`\t`) as the field delimiter in a record.

Compared to my earlier *Twitmographics* prototype, the tab character (`\t`) is favored over other delimiters such as the comma (“`,`”, used in *Twitmographics*), the colon (“`:`”), or the pipe (“`|`”) characters. The rationale is that tabs are not used as a punctuation mark in message text or other Twitter fields, whereas other characters are commonly used as punctuation or field separators within metadata: e.g. the comma is used to separate phrases and also latitude/longitude pairs. Any stray tab characters within fields are sanitized as per the IANA MIME standard *text/tab-separated-values* for tab-separated value files[3].

These output data files are named based on the date/time stamp of its first tweet, and also the unique message ID (Section 4.3.1) allowing its contents to be easily inferred.

Table 5.3 enumerates, in order, the complete set of user and message metadata which are saved to disk in *TweetHarvester* within each tweet record. These fields were discussed in Sections 4.3.1 and 4.3.2; a technical coverage of these metadata items is located in Appendix 1.

5.2.4 Concluding Notes on Post-processing

I conclude this section on the *TweetHarvester* prototype with a few summary notes about post-processing, and a brief comparison between *TweetHarvester* and my earlier on-demand API based *Twitmographics*.

- **Completeness of user data and reduction of API calls:** As opposed to from *Twitmographics*, I need not have to further access the Twitter API again to query additional user information. This is because the Streaming API returns the corresponding user metadata alongside message metadata.
- **Amount of metadata features compared to *Twitmographics*:** Another key difference is that all the algorithms in Sections 4.6 through 4.8 are implemented (as Perl methods) to process the output files directly; compared to *Twitmographics* which uses only a limited subset of the algorithms.
- **Post-processing versus live processing:** It goes without saying that due to the increase in magnitude of the amount of tweets, and the nature of the Streaming API, the time taken by the post-processing stage is significantly longer than the one in *Twitmographics*.

[3] The Internet Assigned Numbers Authority (IANA) has a definition of the *text/tab-separated-values* MIME media type at: <`http://www.iana.org/assignments/media-types/text/tab-separated-values`>.

Table 5.3: The complete list of metadata — from both the user and message domains — collected by *TweetHarvester*.

Category	**Features**
Message domain: tweet properties	• message text string • message source string • message timestamp string • message in-reply-to username • message in-reply-to user ID • message in-reply-to status • message retweeted flag • message retweet count • message favorited flag • message truncated flag
Message domain: JSON entities	• message geo entities (in JSON) • message coordinates entities (in JSON) • message places entities (experimental, in JSON) • message contributors entities (experimental, in JSON)
User domain: common user properties	• user screen name • user ID • user name • user location string • user URL string • user description string • user account creation timestamp • user friends (outdegree) count • user followers (indegree) count • user status total count • user list count (lists to whom he is added by others) • user favorites count

Category	Features
User domain: profile customization attributes	• user default profile (uncustomized) flag • user default profile picture (uncustomized) flag • user profile background color • user profile background URL • user profile background tiling flag • user profile picture URL • user profile link color • user profile sidebar border color • user profile sidebar fill color • user profile text color • user profile background image flag
User domain: other metadata	• user geo-enabled flag • user translator flag • user language code • user notifications enabled flag • user protected flag • user show inline preference flag • user timezone • user UTC offset • user Twitter Verified status • user entities (experimental, in JSON)

5.3 Experimental Data: The *10-Gigabyte Dataset*

In Section 4.2, based on my research and following up with Twitter Inc through their authorized data providers, I have assessed the feasibility of the Streaming API as a data source for large-scale research and analysis. In Section 4.5, I have merely introduced the *10-Gigabyte Dataset* in brief as the choice of training and test data for all my metrics and inference algorithms in Sections 4.6, 4.7, and 4.8.

Within this section, I shall thoroughly document the *10-Gigabyte Dataset*. To begin with, I have conducted several small-scale trials to identify any potential flaws in the *TweetHarvester* prototype (discussed in Section 5.2 prior). After eliminating any potential flaws, I began an exercise in large-scale data collection in November 2011. The design choices for this are as follows:

- **Sampling frequency**: To account for inter-timezone differences in tweet frequency and volume [Pear Analytics, 2009; Zarrella, 2009], I performed sampling of the Streaming API with *TweetHarvester* on an hourly basis. This is to eliminate bias in timezones, i.e. accounting for day-night cycles across the globe.
- **Experiment duration**: Again, based on the difference in daily tweeting habits [Pear Analytics, 2009; Zarrella, 2009], I decided to account for all variations of tweeting activity throughout a given week. Hence, the data collection was conducted for a period of seven days.
- **Resource constraints**: I created a connection to the Twitter Streaming API at hourly intervals, with a maximum self-imposed limit of 50,000 records (of user/message metadata) harvested hourly, to avoid taxing the Twitter servers. Due to network traffic and Twitter server load, however, the maximum limit in the data gathering is infrequently reached.

At the end of the data collection, I removed several records which were corrupted, due to erroneous data supplied by the user causing the record to be unsuitable for data analysis. This included invisible tab characters in a record, generated as artifacts from poorly-written Twitter clients.

The final version of this dataset is codenamed the *10-Gigabyte Dataset*, and will be named as such throughout the rest of this thesis. The following is a summary of the *10-Gigabyte Dataset*:

- **Date range**: Friday 11/11/2011 19:00 — Friday 18/11/2011 19:00 inclusive.
- **Total hours surveyed**: 169 hours (with one tab-separated data file generated per hour).
- **Average message count per hour**: 46,530 records/hour
(minimum = 30,977; maximum = 48,678; standard deviation σ = 3,392).
- **Total message records (after sanitization)**: 7,863,650 records
(each record contains message metadata and its author's user metadata embedded within).

- **Total unique user records (after sanitization)**: 4,491,022 unique users
- **Total storage**: approximately 10.5 gigabytes, uncompressed.

This dataset plays an important role for inference algorithms and approaches which work on empirical real-world Twitter metadata, and will be the subject of subsequent sections in the current chapter: Sections 5.4, 5.5, and 5.6.

5.4 Large-Scale Analysis: Real-life Demographic Properties

Armed with the *10-Gigabyte Dataset*, and my demographic inference algorithms from Section 4.6, this section details my contribution to the body of knowledge with regards to large-scale Twitter demographic analysis.

The inference algorithms from Section 4.6 were run on the entirety of the *10-Gigabyte Dataset* to see how well my algorithms scale on a large real-world dataset, as work on Twitter datasets of such a magnitude are rare in extant literature. Also, the experimental results will be used to reveal any hidden patterns in real-world data embodied by the users on Twitter.

5.4.1 Gender

I ran the *GenderFromName* gender inference algorithm (Algorithm 4.1) — which I had introduced earlier in Section 4.6.1 — on the *10-Gigabyte Dataset* to obtain a gender distribution of real-world Twitter users (Experiment 5.1). This experiment is also published as [Cheong et al., 2012b].

Experiment 5.1. *Large-scale investigation on the gender distribution in the 10-Gigabyte Dataset, using the GenderFromName algorithm [Cheong et al., 2012b].*

Method: From the 4,491,022 unique users in the *10-Gigabyte Dataset*, I have removed 574,175 blank or invalid entries, such as empty strings, non-alphabetic strings, and non-ASCII strings. Algorithm 4.1 is initialized with the SSA dataset as training data [U.S. Social Security Administration, 2011] as it was found to be more up-to-date and produces more accurate results. The algorithm is then used to infer the gender based on first names in the *10-Gigabyte Dataset*.

Results and Discussion: Table 5.4 summarizes the gender distribution on the sanitized 3,916,847 user records in the *10-Gigabyte Dataset* [Cheong et al., 2012b].

The algorithm successfully classified over 2.4 million names (approximately ~61%) into male and female genders. As documented in Section 4.6, good results were obtained in the *10-Gigabyte Dataset*, even for non-Western names from Arabic, Chinese, and Japanese origins [Cheong et al., 2012b]. The remaining 39% could not be classified to either gender due to reasons such as usage of nicknames, non-proper names, and names from other countries which are not in the US SSA dataset [U.S. Social Security Administration, 2011] used for training.

Gender	**Count**	**Percentage**
Male ♂	1,070,490	27.33%
Female ♀	1,319,961	33.70%
Unassigned	1,526,396	38.97%
Total	**3,916,847**	**100%**

Table 5.4: Gender detection: distribution of genders in the *10-Gigabyte Dataset*, as published in [Cheong et al., 2012b]. *Unassigned* indicates that gender could not be inferred from the first name string.

The slightly higher proportion of females to males (viz. ~33.70% versus ~27.33%) corroborates independent real-world surveys on the gender distribution of Twitter users. This comparison is based on publicly-available surveys [Smith and Brenner, 2012; Abraham et al., 2010], which were conducted using phone interviews, Web usage tracking, and similar 'traditional' market research methods.

Based on the aforementioned observations, in essence, the *GenderFromName* algorithm is a highly-scalable and efficient method for determining the genders of Twitter users. Large-scale analysis with *GenderFromName* has the added benefits of identifying potential non-human users (labeled as '*Unassigned*' in Table 5.4), and virtually zero cost, compared to traditional surveying methods.

5.4.2 Geographic Location

As geographic location is another important socio-demographic property, in the same spirit as the previous experiment, I sought out to investigate the real-world geographical distribution of the users in the *10-Gigabyte Dataset*. This investigation is documented henceforth in Experiment 5.2.

Experiment 5.2. *Large-scale investigation on users' geographic distribution in the 10-Gigabyte Dataset, using the Two-phase Hybrid Geocoding algorithm [Cheong et al., 2012b].*

Method: *Two-phase Hybrid Geocoding* (comprising Algorithm 4.2 and 4.3) as per Section 4.6.2) is applied on the 7,863,650 messages in the *10-Gigabyte Dataset* [Cheong et al., 2012b]. This is done on a per-message rather than per-user basis because a user could potentially compose multiple tweets from different geographic locations.

Firstly, I iterated through all the message metadata records in the *10-Gigabyte Dataset*; for each message, I also check for the author's corresponding user metadata. After removing records with no form of location data whatsoever, the *Two-phase Hybrid Geocoding* algorithm was applied on the remainder of the messages.

For each of the records with useful location information, *Two-phase Hybrid Geocoding* will annotate the record with its country of origin. From that, I could keep a tally of the number of tweets found in each country of the world, to allow visualization using a *geographic heat-map*: a world map where color intensities of each country reflect the frequency of tweets originating from within.

Results and Discussion: Over 60.59% of messages (4,764,343 messages) contain no location data whatsoever, and hence unusable for geocoding.

The remaining 3,099,307 messages contain some form of location data: ~36.48% (2,868,649 messages) containing a free-form text string that needs to be pre-parsed with *Geodict* [Warden, 2011], and the remaining ~2.93% (230,658 messages) containing some form of accurate geographic coordinates that can directly be decoded using *CoordinateReverseGeolocation*. Figure 5.3 illustrates how the location data is distributed (free-form location string versus geographic coordinates) across those messages.

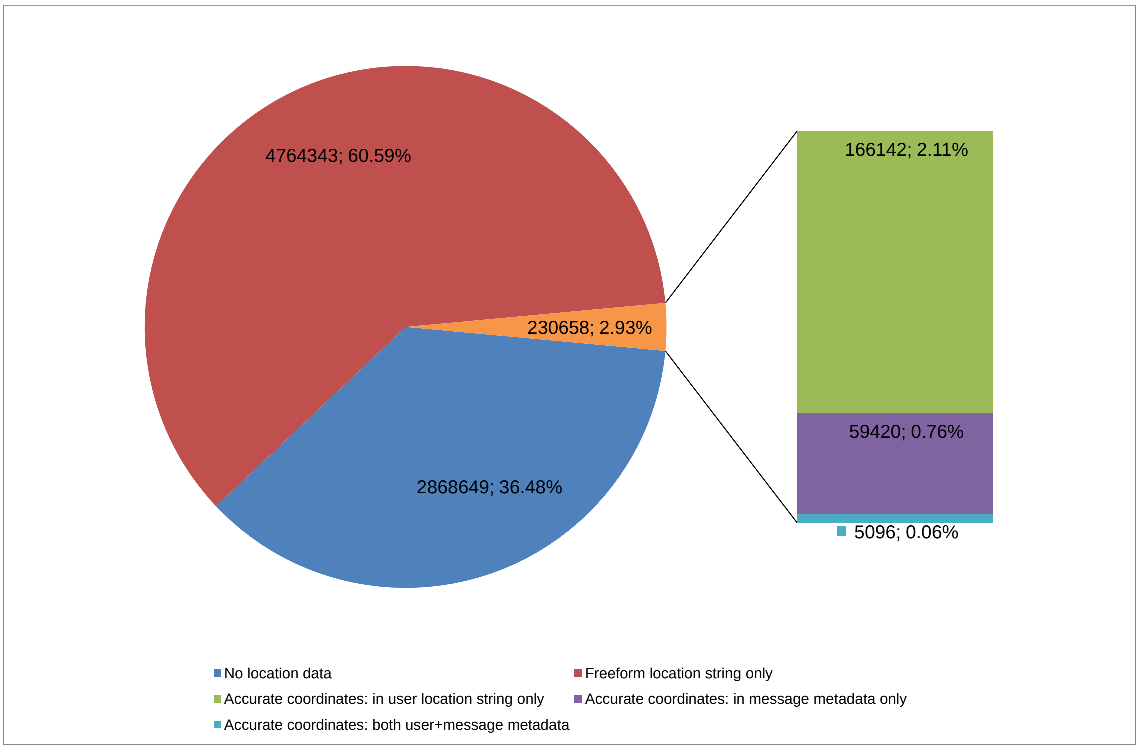


Figure 5.3: The various ways location information can be present in user/message metadata in the *10-Gigabyte Dataset*. As the embedding of *accurate geographic coordinates* in metadata can be done in several different ways, its slice (2.93%) is further sub-divided into its component parts.

After applying *Two-phase Hybrid Geocoding* on the data, I was able to successfully geocode 215,933 accurate coordinates directly. For freeform location text, I successfully parsed and mapped 1,260,910 valid location strings to countries. Figure 5.4 illustrates the results as illustrated on a geographic heat-map, generated with the *OpenHeatMap* service [Warden, 2012]. The color intensities described in the map legend illustrate the number of Twitter messages per country, scaled linearly. This frequency lies in the range of [4 — 25532] — yellow represents the minimum value of 4 users per country, with a linear transition to blue which represents the maximum value of 25,532 users per country.

This experiment on the *10-Gigabyte Dataset* is, to my knowledge, the first time large-scale geocoding has been applied to Twitter metadata in the order of millions of records [Cheong et al., 2012b]. Also, both user (free-form location string) and message metadata

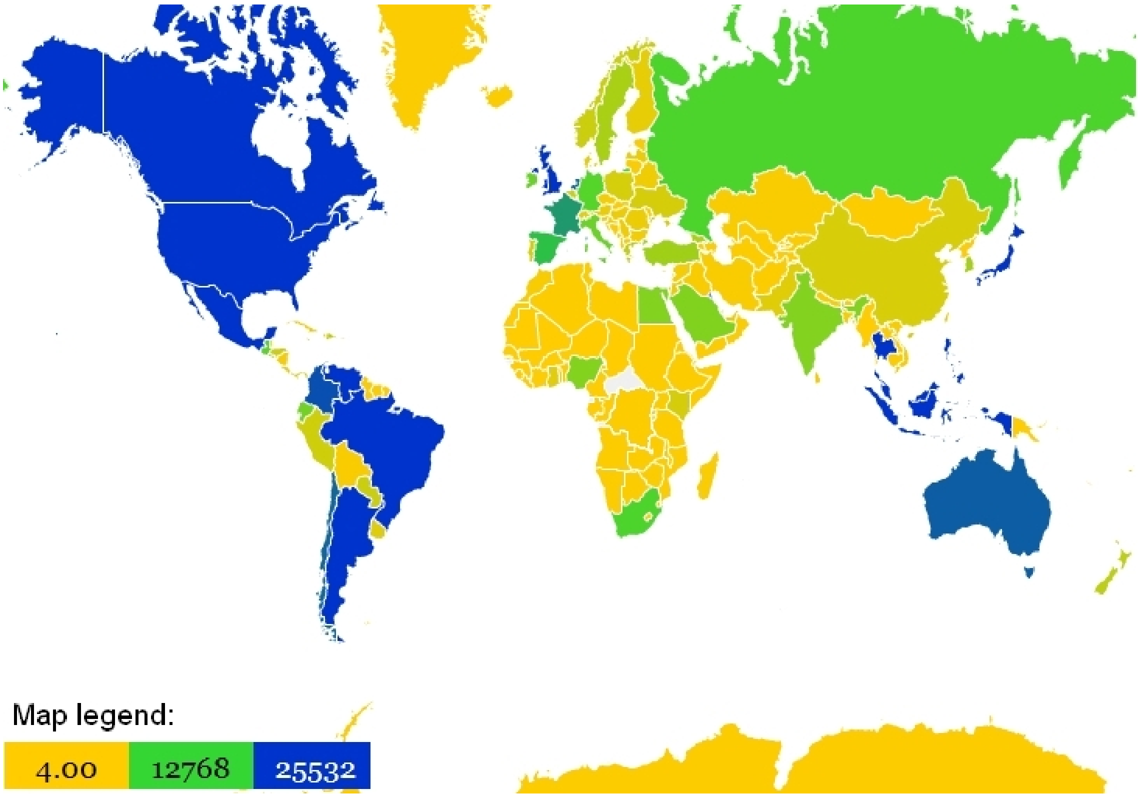


Figure 5.4: Geographic heat-map of Twitter message distribution of the records in the *10-Gigabyte Dataset* [Cheong et al., 2012b]. The color intensities represent the number of messages per country (scaled linearly) as per the map legend. Generated with *OpenHeatMap* [Warden, 2012].

(accurate per-message geographic coordinates) are used in tandem to determine which country a tweet was composed in, which is very rarely done in extant research.

In Experiment 5.2, I have also demonstrated the scalability of my *Two-phase Hybrid Geocoding* algorithm. The millions of tweets in Experiment 5.2 with free-form location text were processed using *Geodict* [Warden, 2011] (the first phase of *Two-phase Hybrid Geocoding*, as per Section 4.6.2) on the *Amazon EC2* cloud computing platform [Cheong et al., 2012b]. An *Amazon EC2 m1.large* instance — with the following resources: 7.5 GB memory, 4 EC2 compute units made up of 2 virtual cores × 2 EC2 Compute Units)[4] — was the underlying platform on which *Geodict* [Warden, 2011] processed the free-form location strings.

Resource-wise, the processing of 2,868,649 free-form location strings in the *10-Gigabyte Dataset* using the *Geodict* algorithm [Warden, 2011] took 5 hours and 53 minutes. At the price of USD$0.604 per hour[5] at time of writing, this merely amounted to USD$3.55. The allocated cloud computing resources (ergo the cost) can 'elastically' be reallocated to cope with the scale of input data, which was hitherto impossible to achieve using commercial third-party services.

[4] Amazon EC2 Instance Types: <http://aws.amazon.com/ec2/instance-types/>.

[5] Figure current as of June 2012. Latest pricing information is available from <http://aws.amazon.com/ec2/pricing/>.

In total, the cost for processing data of this magnitude (with *Two-phase Hybrid Geocoding*) was staggeringly low, contrasting with other commercial third-party services (as suggested in Section 4.6.2):

- The minimum subscription cost for the Google Maps API for Business "...[starts] at just [USD] $10,000 per year"[6].
- The GeoNames premium web service will cost approximately €120 [7] (approximately USD $160, using currency conversion rates as of June 2012) to process ~3 million locations. There are still major caveats, a hard limit of 500,000 lookups per week, requiring almost six weeks to finish processing the *10-Gigabyte Dataset*, and only at a 99% level of service availability.

5.4.3 `source`: Device Classes, Localization, and Mobility

From the investigations on gender and geographic distribution, I proceed to now analyze another demographic inference from the *10-Gigabyte Dataset*: the `source` metadata and its broader implications on demographic knowledge discovery.

In Experiment 4.8, I devised a device classification scheme based on `source` strings, using the *10-Gigabyte Dataset* as empirical evidence, backed with my earlier research (Section 4.6.3, and published as [Cheong and Lee, 2009, 2010c]). Naturally, this section contains a discussion on the empirical distribution of `source` strings in greater depth, in the form of Experiments 5.3 and 5.4.

Experiment 5.3. *Large-scale investigation on modeling the distribution of raw* `source` *strings in the 10-Gigabyte Dataset, as a follow-up from Experiment 4.8.*

Method: In the process of creating the classification scheme, a frequency distribution was created to determine the most frequently-occurring `source` strings in the 7,863,650-message *10-Gigabyte Dataset*; as mentioned in Section 4.6.3.

The first 300 entries in the distribution have been provided in Appendix B, for reference. To visualize the distribution of the `source` strings, I simply plot the frequencies of all 29,098 unique `source`s using a logarithmic scale.

Results and Discussion: The distribution of `source` strings in the *10-Gigabyte Dataset* are as illustrated in Figure 5.5. By using a logarithmic plot on the y-axis (log base 10), the distribution of software clients seem to fit a power law curve.

By assigning each software `source` string a ranking (starting with one for the most-frequently observed, in descending order), I obtained a power-law trendline that can model the experimental distribution, with the equation:

$$Y = 1000000X^{-1.364} \tag{5.1}$$

[6] Google Maps API for Business — Google Earth and Maps Enterprise FAQ, current as of June 2012: <http://www.google.com/enterprise/earthmaps/maps-faq.html>.

[7] From the GeoNames Webservice documentation on Credits; all figures current as of June 2012. GeoNames costs 1 credit per request <http://www.geonames.org/export/credits.html>, with a minimum block of 5 million credits <http://www.geonames.org/commercial-webservices.html>.

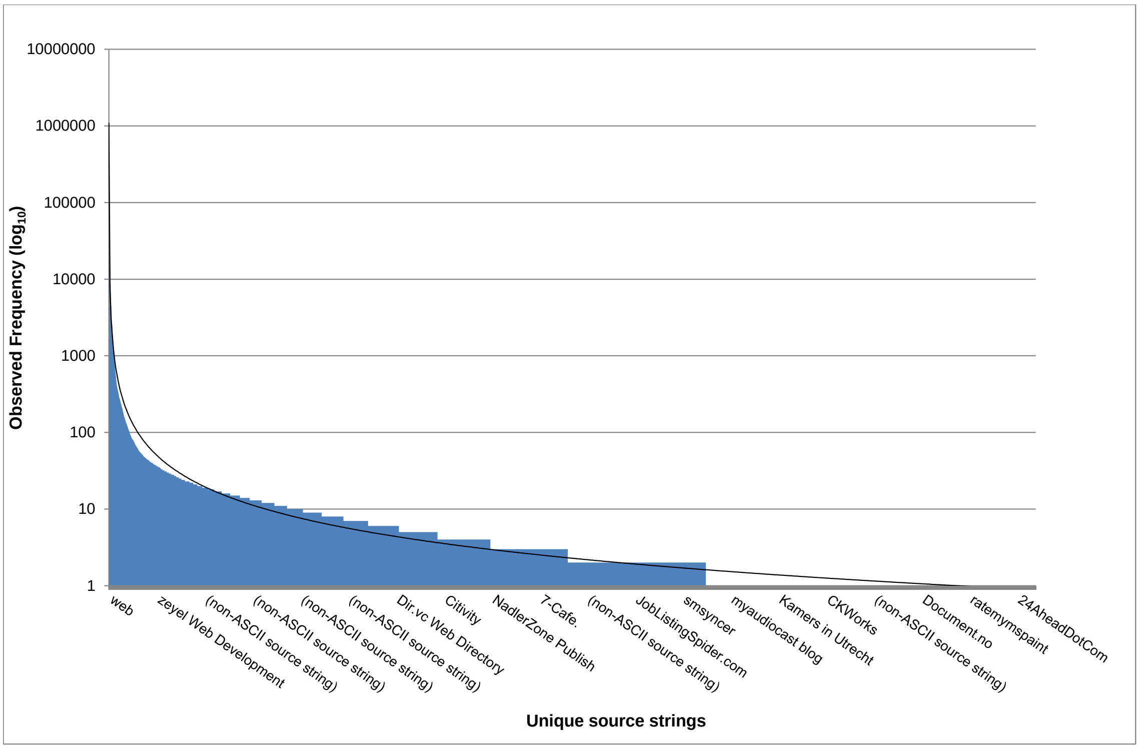


Figure 5.5: A histogram of 29,098 unique Twitter `source` strings as a logarithmic plot, obtained from the *10-Gigabyte Dataset*.

where X is the ranking of a given unique `source` string in terms of frequency, and Y for the observed frequency of said `source` string (the vertical axis in Figure 5.5). For multiple clients with the same frequency, ties are arbitrarily broken.

To determine the closeness-of-fit between the model and my empirical distribution, I calculate the Pearson R^2 coefficient of best-fit [Pearson, 1895], defined as:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \tag{5.2}$$

where n is the number of data points, X is the original data set and Y representing the trendline's data points. The result, Pearson $R^2 = 0.9740$, indicates a close fit between the empirical distribution of the `source` strings in the *10-Gigabyte Dataset* with the estimated power-law model.

Also, to test for monotonicity in the relationship between a source string's frequency and the estimate given using my fitted power-low model, the Spearman Rank-Order Correlation coefficient [Spearman, 1904] is determined. Spearman's rho (ρ) — as it is also called — is defined as:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{5.3}$$

with x being the ranked variables from the original data set, and y the ranked variables from the trendline's data points.

A Spearman's rho of $\rho = 0.9750$ is obtained from the two variables: i.e. the empirical distribution versus the estimated power-law model. A high Spearman's rho suggests the fact that the two variables are monotonically increasing [Spearman, 1904].

To generalize the findings from the is experiment, I have normalized the power-law model equation with respect to the total number of records (7,863,690) in the *10-Gigabyte Dataset*. A general model of

$$P = 12.7167U^{-1.364} \tag{5.4}$$

allows for the estimation of the expected percentage of observations for a particular software client `source` string (P), if its overall usage ranking on Twitter (U) is known.

To explain the distribution of the `source` strings which follows a power law [Barabási and Albert, 1999], I draw upon findings from existing literature on consumer choice and product selection. Consumers of a given type of product tend to exhibit product selection based on Zipf's law (i.e. the power law). This behavior has been documented in prior literature [Brynjolfsson et al., 2006; Goldstein and Goldstein, 2006; ABI Research, 2012], with regards to consumer products ranging from books (on Amazon.com), to movies (on Netflix), to mobile applications (e.g. via Apple's App Store and Google's Play Store). By extension, Twitter users (consumers) exhibit the same behavior when it comes to using ('consuming') Twitter client software.

The results from this larger-scale study does prove that rapid development of new Twitter software has led to an influx of new Twitter clients from all platforms, and subsequently, the increase in `source` strings. Experiment 4.8, conducted in early 2012 (as part of [Cheong et al., 2012b]), revealed that the top 114 (out of approximately 30,000 sources) now account for over 95% of all tweets. Contrast this with the results by Krishnamurthy et al. [2008], three years prior to my work, which found that the "...top dozen sources [account] for over 95% of all tweets" [Krishnamurthy et al., 2008].

Experiment 5.4. *Large-scale investigation on the categories of* `source` *strings in the 10-Gigabyte Dataset using Algorithm 4.8; for the discovery of trends in device classes, mobility, and localization [Cheong et al., 2012b].*

Method: In this experiment, I annotated each message record in the *10-Gigabyte Dataset* with the appropriate category, based on its `source` string. The categorization scheme is listed in Table 4.9. A pie chart is constructed to visualize the distribution of categories throughout the 7,863,690 tweets in the *10-Gigabyte Dataset*.

Results and Discussion: A selection of the top 300 annotated records are given in Appendix B, for reference.

Figure 5.6 summarizes the breakdown of every *10-Gigabyte Dataset* message record into their respective device classes [Cheong et al., 2012b].

As earlier documented in Section 4.6.3 (Experiment 4.8), one can deduce several additional demographic influences based on the *device class* obtained from the `source` metadata item. To recap, these included:

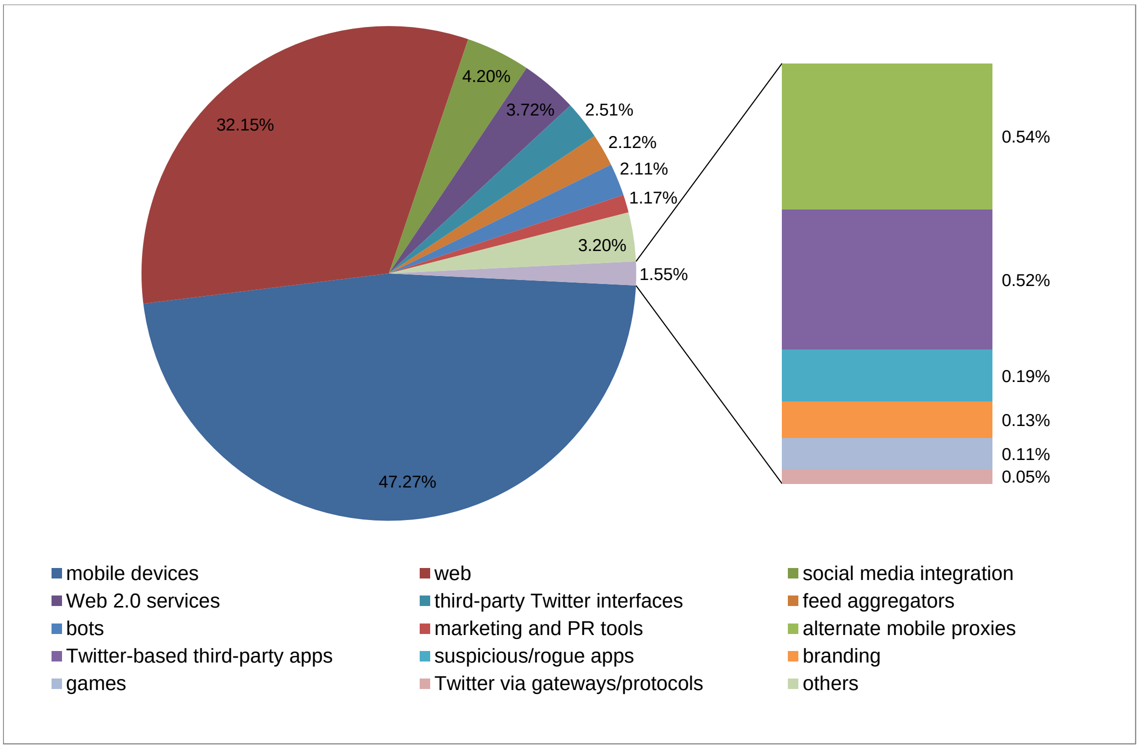


Figure 5.6: Distribution of Twitter client `source` strings in the *10-Gigabyte Dataset* by device class [Cheong et al., 2012b].

- **Deducing user mobility:** a user is more likely to be 'on the move' if he/she is using a *mobile* client as opposed to, say, the web interface on a computer (more in Section 7.2.10 and [Cheong and Lee, 2011]).

- **Detecting censorship:** disproportionate absence of particular device classes (e.g. *mobile* and *social media integrated* clients) during mass events can be used as an indicator of censorship.

- **Determining non-human users and marketing tweets:** frequent appearances of, say, *marketing* and *bot* device classes in a given user's tweets can suggest the fact that the account belongs to a corporation or organization.

From Figure 5.6, a vast majority of users tweet from mobile devices, corroborating existing market studies [Smith and Brenner, 2012; The Nielsen Company, 2010]. This further suggests that a large-scale exploratory exercise (such as Experiment 5.4 on the *10-Gigabyte Dataset* [Cheong et al., 2012b]) can be used as a more accurate and cost-effective way of measuring statistics such as ubiquity of mobile devices, compared to traditional surveys. The second-largest user segment in Figure 5.6 comes from the *web* device class, i.e. the traditional Twitter website at <`http://www.twitter.com/`>.

A comment that can be made regarding these findings is that the current usage trend of Twitter (as of time of writing) is in line with Twitter's original design ethos (cf. [Sarno, 2009], in Section 2.2), which has a strong emphasis on tweeting via a mobile device.

Also, such a large-scale exploration in the veins of Experiment 5.4 [Cheong et al., 2012b] can result in serendipitous discoveries. In the case of the *10-Gigabyte Dataset*, these consist of hidden geographical patterns, hitherto not observed in prior work.

To elaborate, in my observations on the top 300 source strings in the *10-Gigabyte Dataset*, the classification of *bot* programs revealed an interesting pattern. Approximately one-sixth of the surveyed Twitter client software (53 out of 300 unique `source` strings) are targeted towards the Japanese Twitter user base, as they were either exclusively in the Japanese language, partnered or sponsored by Japanese telecommunication providers, or hosted on servers in Japan.

Reviewing the current literature to explain this phenomenon, there are several plausible explanations:

- **Localization**: Terdiman [2008] documented the launch of the Japanese-localized version of Twitter, *Twitter Japan* on April 22, 2008. The Japanese version has a different revenue model (ad-supported), and seeks to promote companies through feeds via *Twitter Japan.* Terdiman [2008] also reported that Tokyo-based tweets are more frequent (almost double) than those from other cities.

- **Japanese user connectivity**: Krishnamurthy et al. [2008] observed that although "Twitter was adopted in Japan later, it has grown quickly to become the third largest region" [Krishnamurthy et al., 2008]. They have also noticed "the more connected nature and popularity of such technologies [as Twitter] in Japan", where Japanese users "in the *.jp* domain... use Japanese[8] [exclusively] to communicate with each other".

- **Niches**: In my earlier case studies Cheong and Lee [2010b,c], I have discovered that Japanese Twitter users also form a significant part in technology-related discussions (cf. Krishnamurthy et al. [2008]'s discussion on technology in Japan). In fact, when performing a clustering experiment (Section 6.3) on tweets discussing the iPhone 3.0 software [Cheong and Lee, 2010b] — a technology-related topic — my algorithm successfully singled out a cluster, purely comprised of Twitter discussion exclusively by Japanese Twitter users [Cheong and Lee, 2010b].

Thus concludes my analyses on real-world *demographic properties* (cf. the methods I first introduced in Section 4.6) with regards to the *10-Gigabyte Dataset*: gender distribution; geographic location and heatmapping; and source string analysis in terms of device classes and mobility. In the subsequent section, I shall explore the *online presence properties* of the users within the *10-Gigabyte Dataset*.

[8] Krishnamurthy et al. [2008] did not exactly document the writing system used in Japanese-language tweets. However, his follow-up work, [Krishnamurthy, 2009] mentioned that "...Japanese users tweet in Kanji". From a cursory examination of Japanese-language tweets on Twitter which I conducted, I observed that such tweets predominantly use Japanese script (*kanji*, *hiragana*, *katakana*) as opposed to Romanized Japanese (*rōmaji*); which corroborates the follow-up by Krishnamurthy [2009].

5.5 Large-Scale Analysis: Twitter Users' Online Presence

Within this section, I will describe several of my contributions in terms of large-scale experiments on user online presence. I experimented with the *10-Gigabyte Dataset* using the algorithms and metrics I earlier proposed in Section 4.7; their results, evaluations, and related discussion are documented herein.

5.5.1 Profile Customization

In the previous chapter, I proposed a novel metric which summarizes a Twitter user's online activity via profile customization. Section 4.7.1 documents my concept of *degree of profile customization* and how it can be used to, among others, determine probability of spam in the message domain and user characteristics.

Experiment 5.5 highlights the application of the *degree of profile customization* concept on the users in a *10-Gigabyte Dataset*. To my knowledge it is the first time such a study involving the Twitter user profile has been conducted on a large scale.

Experiment 5.5. *Large-scale investigation on profile customization scores for users in the 10-Gigabyte Dataset, as a follow-up from Section 4.7.1.*

METHOD: For each of the 4,491,022 unique users in the *10-Gigabyte Dataset*, I calculated the degree of profile customization using Equation 4.1 (Section 4.7.1). To recap, Equation 4.1: *Degree of user customization* = *Avatar presence* + *Profile style customization*.

RESULTS AND DISCUSSION: Table 5.5 documents the distribution of profile customization on the *10-Gigabyte Dataset*, as a co-occurrence distribution of avatar customization (row-wise) with respect to profile customization (column-wise). The total degree of profile customization is listed in parentheses.

	Default profile (no customization)	**Profile customized**
No avatar	13,035 (*degree = 0*)	98,875 (*degree = 1*)
Avatar present	980,707 (*degree = 1*)	3,398,405 (*degree = 2*)

Table 5.5: Co-occurrence distribution of user avatar customization (row-wise) with profile customization (column-wise), for users in the *10-Gigabyte Dataset* (Numbers in parentheses are the total profile customization scores).

From Table 5.5, one can observe the following:

- A majority of Twitter users, 75.67%, customize both their avatar and profile with a *degree of user customization* of **2**.
- The proportion of users with only one form of customization (*degree of user customization* = **1**) is 24.04%.
- Users with no customization to their Twitter profile whatsoever (*degree of user customization* = **0**) constitute only 0.29% of observations.

- The mean degree of profile customization for the whole dataset, $\mu = 1.7538$ (standard deviation, $\sigma = 0.4375$)

These summary statistics by themselves might not provide much insight into the online presence habits of Twitter users. However, the summary statistics obtained across the whole dataset can be used as a good baseline to determine likelihood of spam topics in a given sample of tweets, given user metadata belonging to the tweets' authors; earlier proposed in Section 4.7.1 (Table 4.11).

In fact, from the statistics in Table 5.5, one is also able to infer the proportion of Twitter accounts with *human involvement.* Simply put, *human involvement* in a Twitter user account is defined as any activity *initiated by a human user* in the context of a Twitter user account. Examples include composing a tweet, *following* other users, and — in the context of this experiment — customizing a user's profile.

Recall Experiment 4.9 (in Section 4.7.1), where I discovered that profile pictures and profile style customization will be *absent* in accounts purely run by automated Twitter bot software. Therefore, any sort of profile customization detected in a user account — i.e. a *degree of user customization* greater than 0 — would be proof of human involvement. Even for bot-generated accounts, a minimum degree of human intervention is required to customize a Twitter user's online profile [Schafer, 2010; Collins, 2009; Thomas et al., 2011; Motoyama et al., 2011].

From Table 5.5, at least 4,477,987 (99.71%) user accounts can be said to contain human involvement.

5.5.2 Web Presence from Profile URLs

As documented in Section 4.7.2, extant literature has identified that `url` strings can publicly hint a user's identity and online persona of a given Twitter user's online persona [Dearman et al., 2008; Hughes and Palen, 2009; Schrammel et al., 2008]. I have also devised an URL-classification scheme based on empirical observations on common domain names.

Experiments 5.6 and 5.7 contains a large-scale analysis of the distribution of domains in `url` strings in the *10-Gigabyte Dataset*, their classifications, as well as interpretations of these results in respect of documented real-world behavior.

Experiment 5.6. *Large-scale investigation on the distribution of domain names within user `url` strings in the 10-Gigabyte Dataset, as a follow-up from Experiment 4.11.*

METHOD: From the *10-Gigabyte Dataset*, I first remove user records with blank `url` strings. A total of 1,536,729 non-blank, unique user URL strings were usable for analysis out of 4,491,022 total unique users in the *10-Gigabyte Dataset*. Domain names from the `urls` are then extracted. With these, I constructed a histogram to determine the most frequently-occurring domain names in `url` strings.

RESULTS AND DISCUSSION: Figure 5.7 illustrates the histogram of 337,880 domain names, found in the *10-Gigabyte Dataset*'s `url` strings.

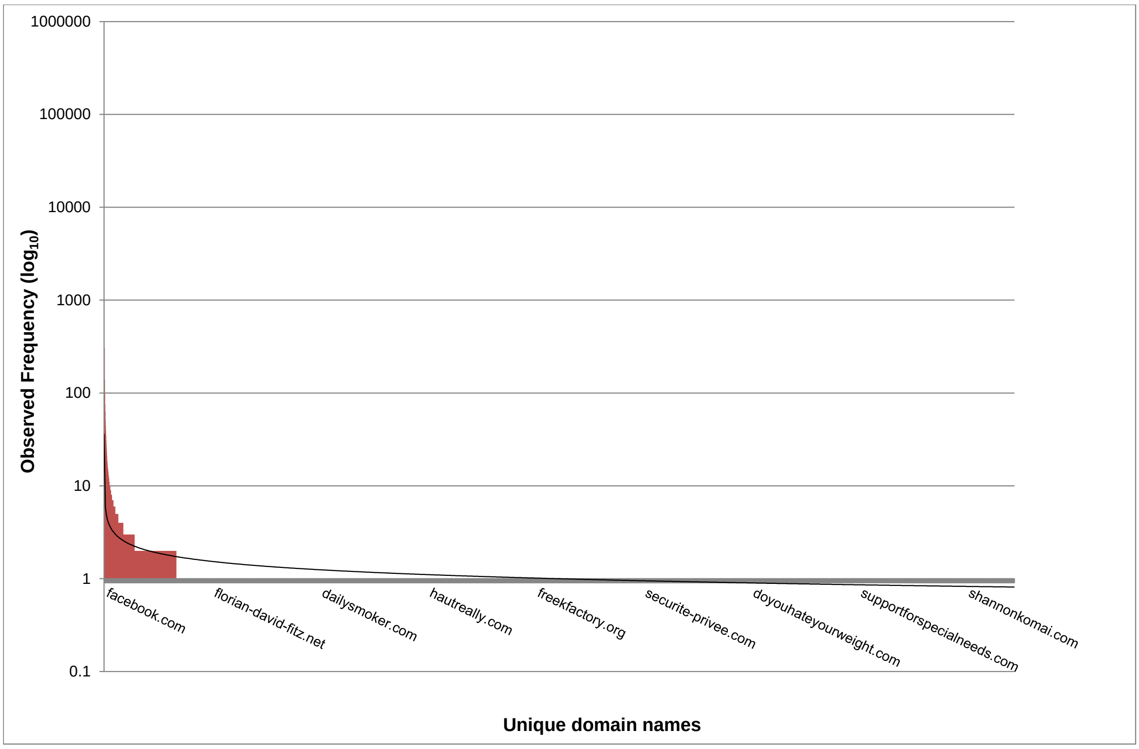


Figure 5.7: A histogram of 337,880 unique domain names found in the 1,536,729 user profile `url` strings as a logarithmic plot, obtained from the *10-Gigabyte Dataset*

By assigning each unique domain name in the distribution with a ranking (starting with one for the most-frequently observed, in descending order), I was able to fit a power-law curve to the experimental distribution, with the equation:

$$Y = 36.111X - 0.298 \tag{5.5}$$

where X is the ranking of a given domain in terms of frequency, and Y for the observed frequency of said domain (the vertical axis in Figure 5.7).

Checking the closeness-of-fit between the power curve with my empirical distribution, I calculated the Pearson R^2 coefficient of best-fit [Pearson, 1895] (Equation 5.2), resulting in a value of $R^2 = 0.5663$.

Similar to the distribution of `source` strings found in Twitter messages (Section 5.4.3), the distribution of unique domain names in user `url` strings follows a power-law.

Compared to `source` strings, however, the frequency distribution in Figure 5.7 exhibits an extremely long tail. Out of a total of 337,880 domain names, this long tail contains 311,102 such names, each appearing only once in the *10-Gigabyte Dataset*. These 311,102 domains, which comprise ~92.07% of the domains, only constitute 20.24% of the total frequency (out of 1,536,729 `url`s).

By removing such outliers — i.e. discarding any domain with a frequency of less than 2 — I was able to fit another power curve which more accurately models my empirical distribution.

In fact, the newly-fitted curve,

$$Y = 6548.4X^{-0.83} \tag{5.6}$$

has a relatively high Pearson R^2 value (Equation 5.2) of 0.9198 (again, with X denoting the ranking of a given domain in terms of frequency; and Y the observed frequency of said domain).

Experiment 5.7. *Large-scale investigation on the types of domains (within `url` strings) in the 10-Gigabyte Dataset, for the discovery of trends in online presences of Twitter users.*

Method: Next, using the classification scheme listed in Table 4.13 (Section 4.7.2), I classify each of the 1,536,729 `url` strings into their appropriate *domain types*. For reference, the first 300 entries in the distribution are listed in Appendix B together with their annotated domain types.

Results and Discussion: Figure 5.8 illustrates the percentage distribution of unique domain types found in the *10-Gigabyte Dataset*.

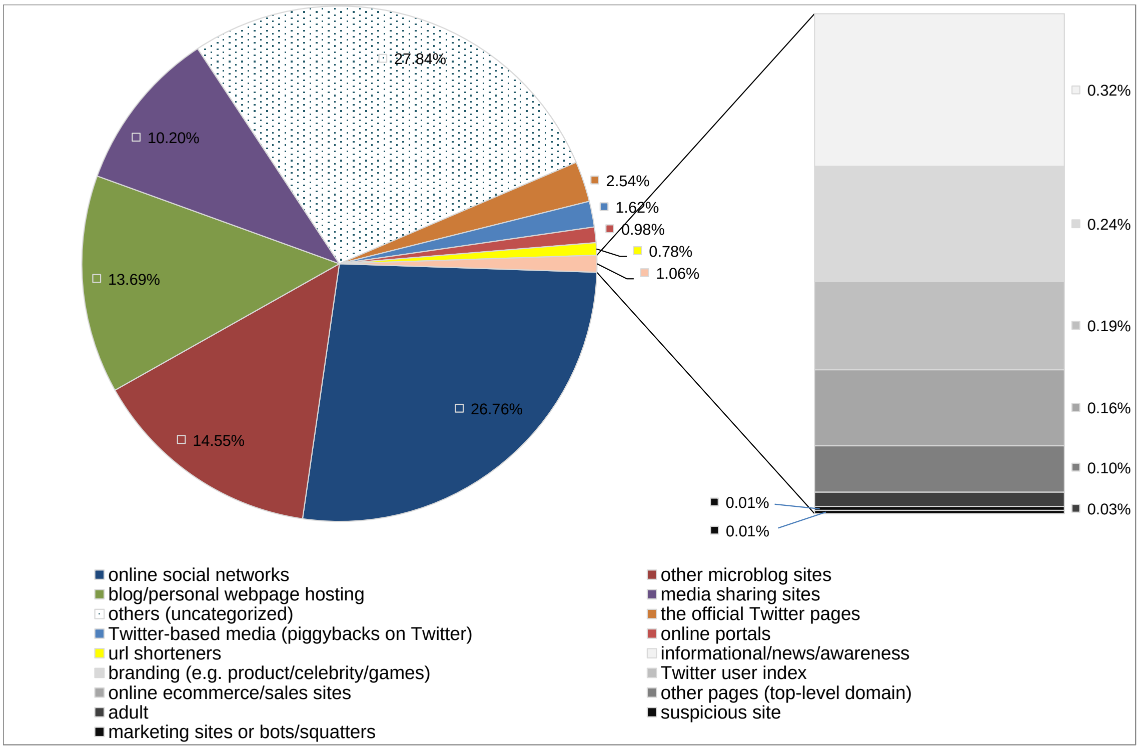


Figure 5.8: Pie chart illustrating the percentage distribution of unique domain types found in `url` metadata, obtained from the *10-Gigabyte Dataset*.

From Figure 5.8, notice that URLs linking to *online social networks*, *other microblog sites*, and *media sharing sites* comprise over half of the `url` strings in the *10-Gigabyte Dataset*. This is seen as a shift to social media as a form of online presence [Schawbel, 2011].

Sites such as Facebook are able to reveal "small slices of our professional life, hobbies or youthful misdeeds... [which can sometimes be] viewed out of context" [Fry, 2008]. It is not unusual for a person to cross-link their presences on another social media platform [Kim, Jeong and Lee, 2010; Ramsden, 2008]. In the context of Twitter, this is made possible by using profile `url` strings to link to their Facebook account, among other methods of Facebook integration.

In a similar vein, a Twitter user also commonly advertises his or her personal web page (~13.69% of the `url` strings found), as such pages serve as a "...home that offers context for [one's] various online activities, building a mosaic out of what would otherwise be baffling fragments" [Schawbel, 2011].

The rest of the distribution in Figure 5.8 is comprised of not-so-commonly-used websites, each serving a niche purpose: from advertising and brand promotion; to sharing informational links within a user's profile *sans* tweets; to aggressive marketing and Search-Engine Optimization (SEO) techniques.

5.5.3 User Connectivity

As discussed in Section 4.7.3, existing studies on Twitter user metadata [Java et al., 2009; Krishnamurthy et al., 2008; Huberman et al., 2008b; Kwak et al., 2010; Makice, 2009b; Russell, 2011a] emphasized the reconstruction and characterization of a Twitter user's social graph. The REST-`user` API suffices to accomplish this task: the follower and friend network of a particular user can be traversed or *crawled* using API commands such as `GET followers/ids` and `GET friends/ids` respectively [Twitter Inc., 2012a]. In turn, the followers and/or friends of these users can then be traversed, bound only by Twitter API rate limits, network resources, and computing time.

Using this aforementioned simple methodology as a basis, authors of several studies [Kwak et al., 2010; Krishnamurthy et al., 2008; Java et al., 2009] were able to study a Twitter user's network topology and characterization, identify opinion 'leaders', and summarize the overall Twitter user distribution. There are also many related case studies which cover summary statistics on friends and followers, such as the distribution of a user's friend count with respect to the number of followers [Pear Analytics, 2009; Zarrella, 2009].

Having said that, the purpose of this section is not to replicate these aforementioned studies. Rather, I focus on the latent characteristics of the 4,491,022 unique users in the real-world *10-Gigabyte Dataset*. Since the *10-Gigabyte Dataset* was sourced from the Streaming API, only summary statistics are available for a given user in terms of his or her social graph — `followers_count` and `friends_count` — as earlier alluded to in Section 4.2. Hence, a viable method of analysis of social graph connections for the set of unique users comprising the *10-Gigabyte Dataset* is the examination of the follower/friend ratio (FFR) summary statistic (Section 4.7.3).

This examination will be conducted in two parts. Firstly, Experiment 5.8 replicates two extant studies on the FFR on the *10-Gigabyte Dataset* as a basis for comparing the connectivity of users in my *10-Gigabyte Dataset* with the results found in existing research. In doing so, I also illustrate that the FFR is a viable metric for the large-scale study of users, in the absence of detailed user graph data caused by the abovementioned Twitter API limitations.

Experiment 5.8. *Comparing the FFR characteristics of the 4,491,022 unique users in the overall 10-Gigabyte Dataset with those found in current literature.*

METHOD: For this experiment, I refer to two existing experimental trials which can be replicated on my *10-Gigabyte Dataset*:

Trial I: Schafer's 2010 study [Schafer, 2010] on anomalous Twitter accounts which included a thorough analysis of 66,250,639 user 'nodes' found in Twitter Social Graph (conducted before Twitter's rate-limiting of the REST-`user` API, cf. Section 4.1.3). In this study, Schafer generated a histogram of the *friend-follower ratio* — the *inverse* of my FFR metric in Equation 4.2 (Section 4.7.3) — to better understand the distribution of user graph links on Twitter. Replicating this experiment merely involves obtaining the *inverse* of my FFR (yielding the ratio of friends to followers) for each of the users in my *10-Gigabyte Dataset*, and comparing the resulting frequency distribution (in $\log_2$ scale) with the one published in [Schafer, 2010]

Trial II: The pioneering exploratory study by Java et al. [2009] analyzed the follower/friend distribution of 87,897 user nodes. The data from [Java et al., 2009], presented as a scatterplot, involves not only the simplified FFR, but also the actual counts of followers (or friends). For the purposes of comparison, I decide to replicate Java et al. [2009]'s experiment on my *10-Gigabyte Dataset*: by decomposing the FFR into the its elements of `followers_count` and `friends_count`; these variables are then illustrated for each user on a scatterplot. The correlation coefficient, which was used to determine the "...correlation between the indegree and outdegree for Twitter users" [Java et al., 2009] is also calculated from the generated scatterplot and used as basis for comparison. Java et al. [2009] did not elaborate on the method used, which could be either the Spearman Rank-Order correlation coefficient (ρ, defined in 5.3) or the Pearson product-moment correlation coefficient (r, defined in 5.2) depending on context. For completeness, I have included both these coefficients in my analysis.

RESULTS AND DISCUSSION: From replicating the two selected experimental trials on the *10-Gigabyte Dataset*, the following results were obtained:

Trial I: Replicating Schafer's 2010 study [Schafer, 2010] on the *10-Gigabyte Dataset*, I obtain a result close to that of Schafer [2010] (Figure 5.9).

Similar to the original results by Schafer [2010], the peak of both graphs are "near 0" [Schafer, 2010]. For both graphs in Figure 5.9, "...the left side of the graph (more followers than friends) has a much shallower drop-off than the right side of the graph (more friends [than] followers)" [Schafer, 2010]. The only difference between the two graphs is the sharper peak (Figure 5.9, right), from the *10-Gigabyte Dataset*. This is due to the more spread-out distribution of the FFR (and its inverse, as demonstrated in this study); reflecting a shift away from the 1:1 relationship between the number of friends and followers evident in earlier years.

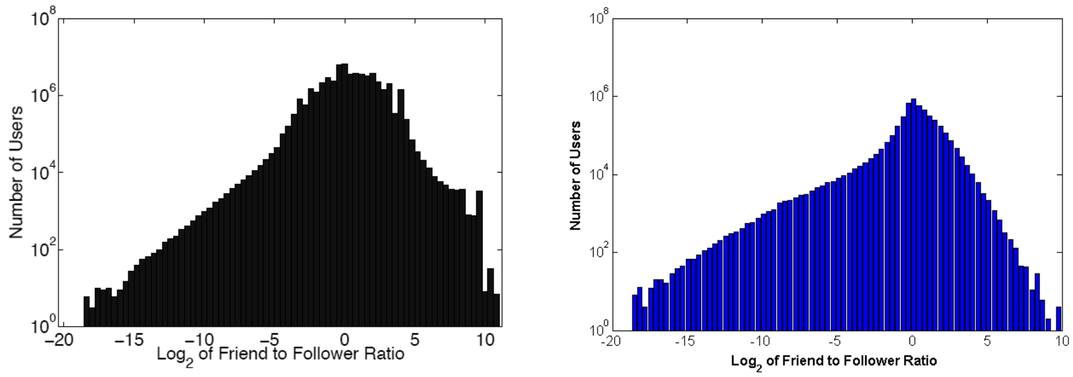


Figure 5.9: Results of the original experiment by Schafer [2010] (left, in grayscale). Replication of his experiment on the *10-Gigabyte Dataset* (Trial I) resulted in a similar histogram (right, in color).

Trial II: Replicating Java's 2007 study [Java et al., 2009] on the *10-Gigabyte Dataset*, I observe a different trend in the absolute counts of followers to friends (Figure 5.10); despite the fact that FFR distribution has a fairly consistent pattern due to *Trial I* of this experiment.

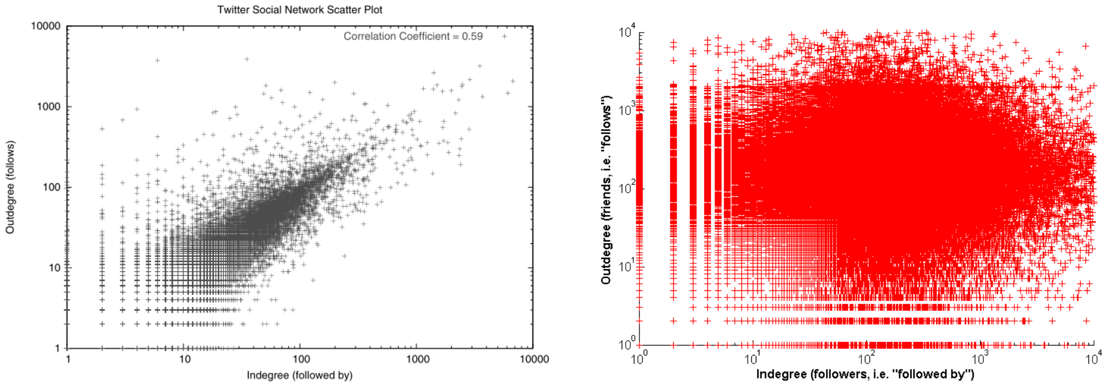


Figure 5.10: Results of the original experiment by Java et al. [2009] (left, in grayscale). Replication of his experiment on the *10-Gigabyte Dataset* (Trial II) resulted in a distinctly more spread-out scatterplot (right, in color).

Both the scatterplots in Figure 5.10 are drawn on the same axes and scales. The color plot (Figure 5.10, right, in red) reveals a highly different distribution of absolute friend and follower counts in the *10-Gigabyte Dataset*, compared to the original experiment by Java et al. [2009] (Figure 5.10, left, in grayscale). This is caused by the evolution of the *following* patterns on Twitter since the results by Java et al. [2009] were published (approximately five years, at time of writing this thesis).

In terms of the correlation coefficient, for Java et al. [2009] (Figure 5.10, left), the obtained "correlation coefficient" was 0.59, signifying "...that users who are followed by many people also have large number of friends" [Java et al., 2009]. However, the authors did not specify the exact correlation coefficient used. From my replication of the experiment on the *10-Gigabyte Dataset*, I calculated both the Spearman

Rank-Order correlation coefficient (Equation 5.3, resulting in ρ=0.7965) and Pearson's product-moment correlation coefficient (Equation 5.2, resulting in r=0.1504). Due to a low Pearson product-moment correlation coefficient, there is a weak linear correlation between the absolute counts of followers and friends. The Spearman Rank-Order correlation coefficient, on the other hand, is a strong indicator of how the number of friends will monotonically increase based on the number of followers, agreeing with the qualitative results of Java et al. [2009].

From both the replicated trials in Experiment 5.8, the FFR is shown to be a reliable indicator on the overall distribution of user connectivity on Twitter (Figure 5.9). Over the years since Java et al. [2009] was published, a significant change in the distribution of absolute friend and follower counts on Twitter can be observed (Figure 5.10). Such a pattern was evident in related observational studies of users Barracuda Networks, Inc. [2010]. However, the relationship between the two variables can still be described using a monotonic function.

My second and final experiment in this section involves characterizing the overall distribution of the Twitter user graph; achieved by examining the overall FFR ratios from the *10-Gigabyte Dataset*. As recorded in Section 3.1's review of existing literature, the connections between users on Twitter follow a power-law (i.e. $y = ax^k$) distribution [Java et al., 2009], if the effects of outliers were disregarded [Kwak et al., 2010]. In fact, Experiment 5.9 resulted in a distribution of *friends-to-followers* (the inverse of the FFR I defined in 4.2) which follows a power-law distribution, corroborating the findings of Schafer [2010]. Hence, in Experiment 5.9, the histogram of FFRs in the *10-Gigabyte Dataset* will be studied and tested for compliance with a power-law model.

Experiment 5.9. *Large-scale investigation on modeling the frequency distribution of follower/friend ratios (FFRs) of users in the 10-Gigabyte Dataset, and to verify that it follows a power-law.*

Method: In order to generate the frequency distribution of FFRs in the *10-Gigabyte Dataset*, the FFRs (viz. Equation 4.2) for each of the 4,491,022 unique users in the *10-Gigabyte Dataset* are calculated. Outliers which are uncharacteristic of the overall distribution are then removed from the distribution. A power curve is fitted to the frequency distribution, and its closeness-of-fit measured using Pearson's R^2.

Results and Discussion: Figure 5.11 illustrates the histogram of FFRs in the *10-Gigabyte Dataset*.

Before constructing the histogram, I have removed 13 outliers at the far end: i.e. users with disproportionately high FFRs of {459879, 479124, 498331, 514170, 521151, 568759.4, 745249, 769993, 926073, 1196705, 1892270, 3099603, 4330976}, to minimize the effect of such extreme outliers on the overall distribution. The properties of outliers in the *10-Gigabyte Dataset* were already documented in Table 4.14 (Section 4.7.3).

The FFR histogram, *sans* outliers, has the following properties.

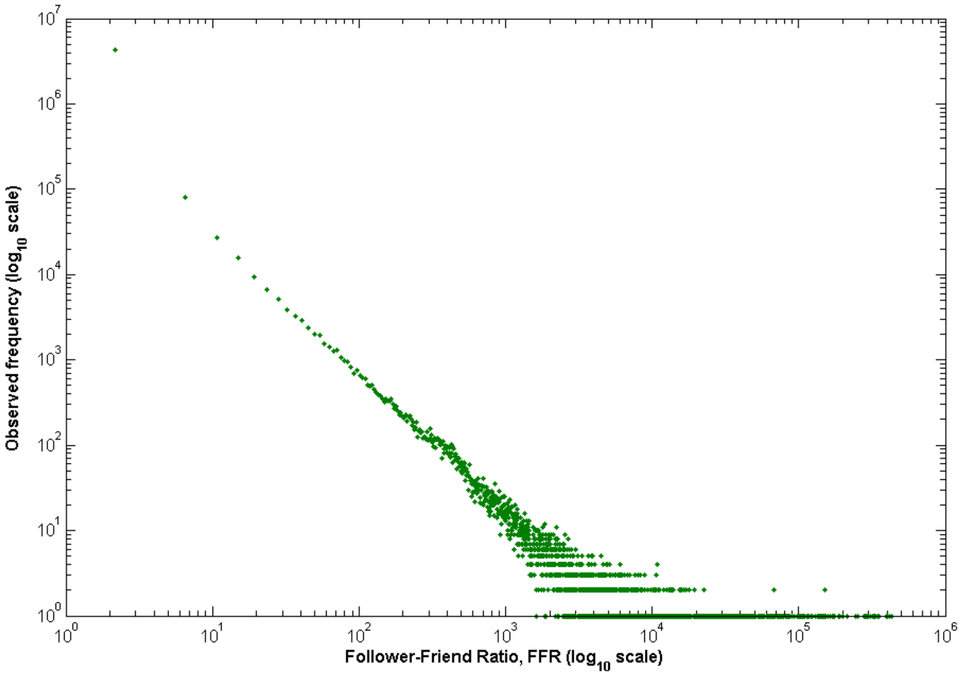


Figure 5.11: Histogram illustrating the FFR distribution of the *10-Gigabyte Dataset*, in a log-log plot.

- Range: 0–431,193
- Median: 0.8696
- Mean, μ: 12.3380 (standard deviation, σ: 989.2536).

To check for compliance with the power-law, I fitted a power curve with the equation $Y = 5900100X^{-1.9002} + 0.93988$ to the histogram (given X = FFR cf. the horizontal axis in Figure 5.11; and Y = observed frequency for the given FFR cf. the vertical axis in Figure 5.11). Using the Pearson R^2 coefficient of best-fit [Pearson, 1895] in Equation 5.2, I obtained a value of $R^2 = 0.94112$ which suggests the closeness-of-fit of the power-curve.

This goes to show that the FFR distribution found in the *10-Gigabyte Dataset*, *sans* outliers, does indeed follow a power-law distribution [Java et al., 2009; Kwak et al., 2010; Schafer, 2010]. This is indicative of the fact that the Twitter user network exhibits *scale-free* properties [Barabási and Albert, 1999], as posited by existing literature seen throughout Chapter 3: e.g. [Krishnamurthy et al., 2008; van Liere, 2010; Stonedahl et al., 2010]; and from Section 7.3 later in this thesis (published as [Cheong et al., 2012a]).

5.5.4 User Loyalty and Usage Frequency

To conclude the study of the online presence of Twitter users as represented by the *10-Gigabyte Dataset*, this section studies two account activity metrics of Twitter users: normalized account age and messaging frequency metrics, first introduced in Section 4.7.4 (Experiments 5.10 and 5.11 respectively).

Experiment 5.10. *Large-scale investigation on the frequency distribution of users' normalized account ages in the 10-Gigabyte Dataset, and its defining characteristics.*

METHOD: As per Section 4.7.4 earlier, the normalized account age for a given user is defined by Equation 4.3 (Account age = Observation date − `created_at` + 1).

I generated a frequency distribution of normalized account ages for each unique user in the *10-Gigabyte Dataset* (4,491,022 in total). I also consulted with existing literature on the growth trends and user base characteristics of Twitter to help interpret the defining characteristics of the constructed histogram.

RESULTS AND DISCUSSION: Figure 5.12 illustrates the frequency distribution of users' normalized account age found in the *10-Gigabyte Dataset*.

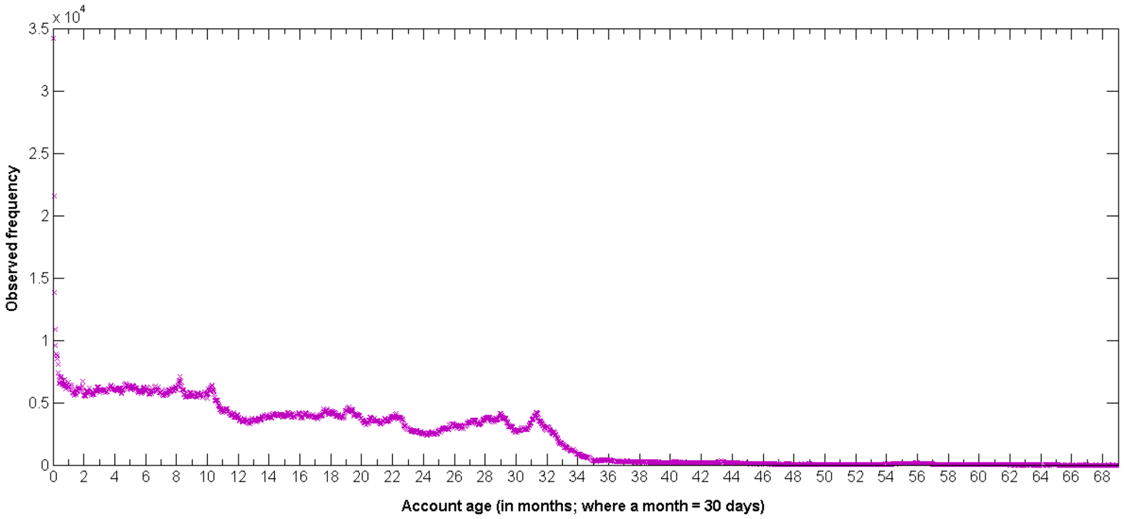


Figure 5.12: Distribution of the normalized account age of unique users in the *10-Gigabyte Dataset*.

Before attempting to explain the features of the obtained distribution, a cursory overview of its statistics is as follows:

- Range of normalized account age: 1–2,068 days
- Median: 386 days
- Mean, μ: 440.9375 days (standard deviation, σ: 316.8955).

At first glance, the overall frequency distribution roughly approximates that of a negative-exponential distribution; which can be modeled by the equation:

$$Y = 8093^{X/-609.23} \tag{5.7}$$

with Y being the observed frequency (represented by the horizontal axis in Figure 5.12), and X being the normalized account age in days (presented in terms of months in Figure 5.12 for brevity). A Pearson $R^2 = 0.8037$ was obtained with this curve. The proposed model takes into consideration the sharp drop in the initial y values as illustrated in Figure 5.12.

If 'new' accounts were removed, the data will more closely fit the negative exponential distribution. By removing users with a normalized account age of a month (30 days) or less, for example, a negative-exponential model closer fits the distribution:

$$Y = 7895.4^{X/-623.49} \tag{5.8}$$

Again, Y denotes the observed frequency and X denotes the normalized account age in days. The Pearson R^2 (Equation 5.2) obtained from the model in Equation 5.8 is 0.8605.

By consulting the literature, and from the negative-exponential model proposed above, there are two main factors which contribute to the unique distribution of normalized account ages of users in the *10-Gigabyte Dataset*:

1. **New users 'testing the waters':** As mentioned, a sharp spike occurs in the number of users with a low normalized account age (the initial portion of the distribution). This spike rapidly drops an order of magnitude in the first few days, evident in a cursory examination of the first few histogram entries: {34177, 21574, 13813, 10851, 9584, 8867, 8716, ...}. There is, however, a simple explanation for this initial spike. Referring to the definition in Section 4.7.4, the normalized account age can simply be summarized as: *the (normalized) number of days since a Twitter user account was created, calculated at the time the user publishes his/her latest tweet.* In terms of the *10-Gigabyte Dataset*, the population of users with a low normalized account age (the initial spike in Figure 5.12) consist of new Twitter accounts, which have composed at least one tweet in the short span of time post-account creation. Thus, the *10-Gigabyte Dataset*, created using the Streaming API (Section 4.5), contains an abundance of tweets from users who have created new Twitter accounts for this very purpose. Existing literature [Heil and Piskorski, 2009; Zarrella, 2009; Pear Analytics, 2009; Sysomos Inc., 2010] documents the fact that users typically show more activity in the early life of their Twitter account; after their first few tweets, a significant proportion of users leave their Twitter accounts idle.

2. **Noteworthy growth spurts of the Twitter user base:** The various spikes in the distribution of users by normalized account age can be attributed to growth spurts of the Twitter user base. For example, Zarrella [2009] documented two spikes in Twitter user growth, in March 2007 (~1716 days or ~57 months before the *10-Gigabyte Dataset*) and March 2008 (~1350 days or ~45 months before the *10-Gigabyte Dataset*). Sysomos Inc. [2010] documented growth spurts of the Twitter user base at around December 2008 (~1075 days or ~36 months before the *10-Gigabyte Dataset*), and at January 2010 (~679 days or ~23 months before the *10-Gigabyte Dataset*); both of these upward growth trends are visible in Figure 5.12. (However, as of time of writing, however, exact quantitative figures for Twitter's user base and/or it's growth in 2011-2012 are yet to be made available).

Experiment 5.11. *Large-scale investigation on modeling the frequency distribution of users' average posts per day in the 10-Gigabyte Dataset.*

Method: In this final experiment under the banner of user loyalty and usage frequency, a frequency distribution of message frequencies for each of the 4,491,022 unique users in the *10-Gigabyte Dataset* will be generated.

As defined in Section 4.7.4, the average message frequency for a given Twitter user is defined as per Equation 4.4 (Message frequency $= \frac{\Sigma\text{number of tweets posted } (\texttt{statuses\_count})}{\text{Account age}}$).

Results and Discussion: Figure 5.13 illustrates the frequency distribution of average posts per day.

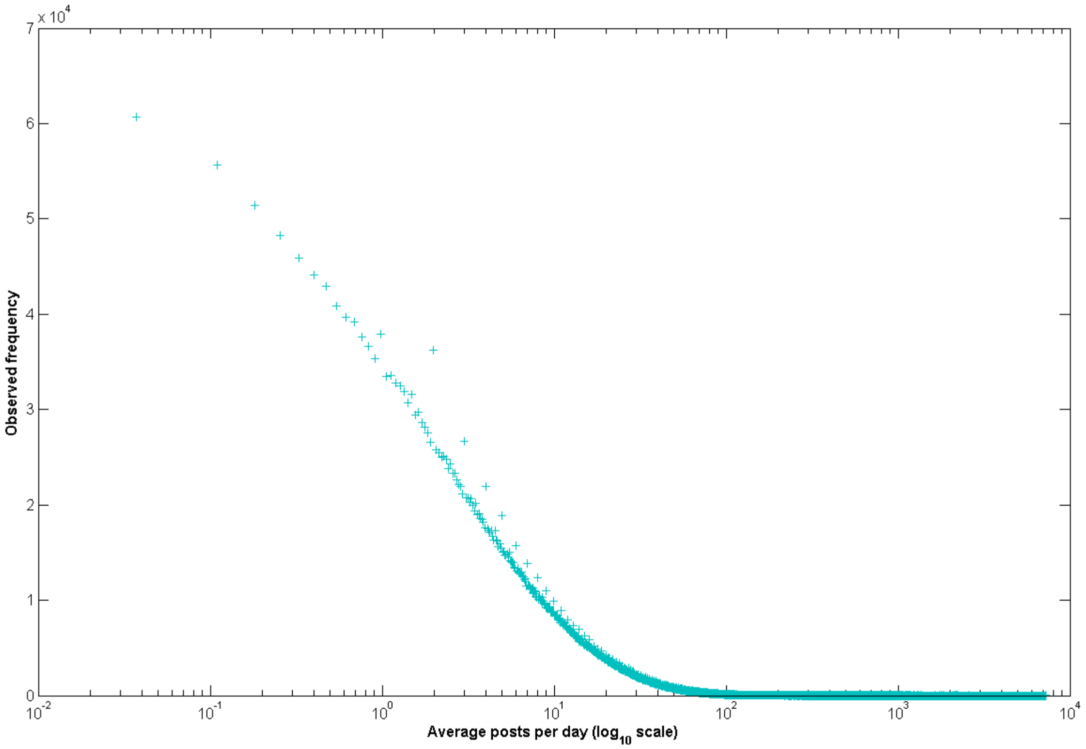


Figure 5.13: Distribution of average posts per day of unique users, as a logarithmic plot, of the *10-Gigabyte Dataset*.

The following summary properties were observed from the histogram shown in Figure 5.13.

- Range of messages/day: ~ 0–7,254 messages/day
- Median: 6.9761 messages/day
- Mean, μ: 15.3038 messages/day (standard deviation, σ: 27.1260).

The frequency distribution of average posts per day from the *10-Gigabyte Dataset* approximates a negative-exponential distribution, akin to the distribution of the normalized account ages of users (Experiment 5.12). Fitting a negative-exponential curve to the data, where X = average posts per day (the horizontal axis in Figure 5.13) and Y = observed frequency of users (the vertical axis in Figure 5.13):

$$Y = 40037X^{-0.1551X} \tag{5.9}$$

a Pearson R^2 coefficient of 0.95632 is obtained (via Equation 5.2). The negative-exponential distribution of average posts per day corroborate with existing patterns in related social media; most prominently seen in the case of Wikipedia [Ratkiewicz et al., 2010; Priedhorsky et al., 2007; Mendoza et al., 2010].

5.6 Large-Scale Analysis: Communication Patterns

The final part of my large-scale empirical study on the *10-Gigabyte Dataset* deals with the summary properties of the messages harvested in the dataset. It is pertinent to note that the focus of this thesis is on pattern discovery and inference generation from metadata, as opposed to textual analysis (which is the sole focus of e.g. [Horn, 2010; Shamma et al., 2009; Jansen et al., 2009b]).

As such, discussion on the summary statistics and latent properties of tweets in my large-scale study as earlier defined in Section 4.8.1 would suffice, as opposed to those studies exclusively dealing with the Twitter message domain.

5.6.1 Distribution of Message Length

Message length, as discussed prior in Section 4.8.1, helps characterize message text in terms of how much information is conveyed in a given tweet. In this section, I shall first generate the distribution of message lengths found in the collection of 7,863,650 messages in the *10-Gigabyte Dataset*, and quantitatively evaluate the obtained distribution. My evaluation will be compared to existing studies [Yoshida et al., 2010; Zarrella, 2009], to determine if the results obtained are consistent and reproducible (Experiment 5.12).

Experiment 5.12. *Large-scale investigation on modeling the distribution of tweet message lengths in the 10-Gigabyte Dataset.*

METHOD: The tweet content (`text` metadata field) for the 7,863,650 messages in the *10-Gigabyte Dataset* is extracted. I excluded tweets containing non-ASCII characters, such as those encoded in Chinese-Japanese-Korean (CJK) or Arabic character sets (cf. prior discussion in Section 4.8.1 and Section 5.2).

Also excluded are messages containing extended special characters, which are encoded as multiple ASCII characters due to the disproportionate number of bits required for encoding. Finally, HTML-escaping is performed on certain messages, where HTML-escaped symbols such as `&lt;` and `&gt;` are converted into the appropriate symbols, i.e. *less-than* (`<`) and *greater-than* (`>`), respectively.

The lengths of all the pre-processed tweets were counted, from which a frequency distribution was created. To summarize the graph, I calculate summary statistics and derive a model for the obtained distribution. For validity, the results obtained from this experiment will also be compared to extant literature [Yoshida et al., 2010; Zarrella, 2009].

Results and Discussion: During pre-processing, 341,215 messages were discarded as they consisted of either empty strings, or entirely of non-ASCII characters. Also, 367 messages which contained a mixture of ASCII and non-ASCII characters were removed. Such extended non-ASCII characters were encoded as multiple ASCII characters due to their bitwise representations; the presence of which artificially skewed the character count past the 140-character limit.

Figure 5.14 illustrates the distribution of message lengths, from the remaining 7,522,068 messages in the dataset. This distribution has a median of 51 characters, and a mean, μ of 59.6268 characters (with a high standard deviation, $\sigma = 40.2325$ characters). Factoring in the removal of 10-20 character URLs by Yoshida et al. [2010], the mean obtained from the current experiment (59.6268 characters) is similar to the mean obtained by Yoshida et al. [2010] which is 41.51 characters for human tweets (i.e. 51.51–61.51 characters when the URLs are factored in).

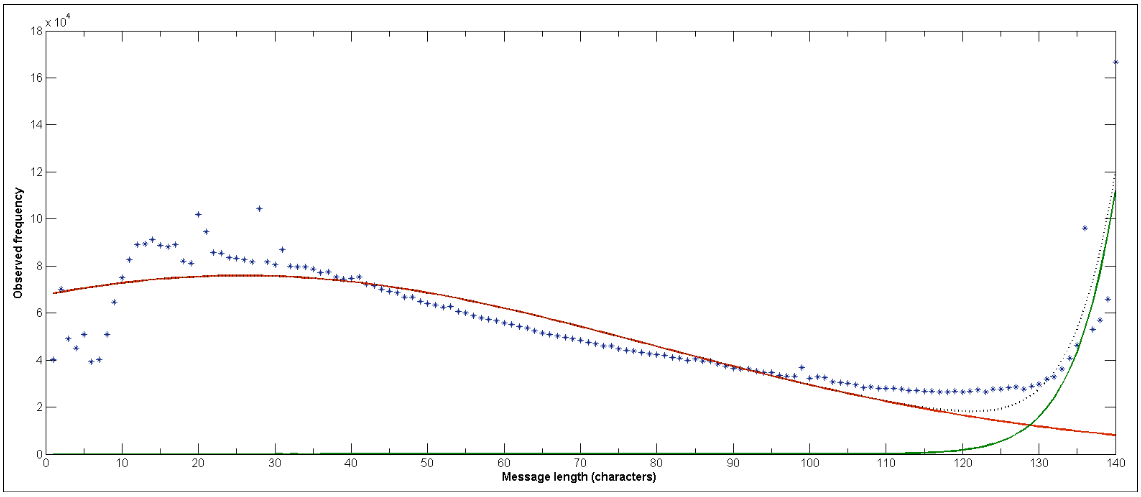


Figure 5.14: Histogram of message lengths, as a logarithmic plot, for 7,522,068 tweets (post-sanitization) found in the *10-Gigabyte Dataset*, which contains a bimodal distribution. The red and green curves are non-normalized Gaussians which model the first and second maxima in the graph respectively. The mixture of both Gaussians is represented as a dotted black curve.

In its entirety, this distribution approximates a polynomial of the sixth degree. However, a polynomial of such a high order will result in a model that overfits the data, and isn't a good choice for modeling the distribution.

The presence of two local maxima in Figure 5.14 's bimodal distribution (at $x = 28$ characters, and $x = 140$ characters, respectively) is indicative of a mixture distribution of two Gaussians. Fitting two non-normalized Gaussian distributions around these two maxima, I was able to model the overall distribution as a combination of two non-normalized Gaussians; firstly (the red curve in Figure 5.14):

$$y(x) = 75899e^{-\frac{(x-25.726)^2}{2\times 53.959^2}} \tag{5.10}$$

and secondly (the green curve in Figure 5.14):

$$y(x) = 52612000e^{-\frac{(x-207.7)^2}{2\times 19.299^2}} \tag{5.11}$$

The combination of the two non-normalized Gaussians is illustrated as a dashed black line in Figure 5.14.

The shape of the distribution in the *10-Gigabyte Dataset* is similar to the results by Yoshida et al. [2010] and Zarrella [2009]: an initial local maximum and the characteristic sharp spike towards the $x = 140$ character limit were present in all cases. In the case of Yoshida et al. [2010], it is pertinent to note that the spike is detected at the 110-120 character mark [Yoshida et al., 2010], as they removed the length of URLs from their overall tweet length.

The two maxima can be attributed to:

1. **The initial local maximum centered at 28 characters:** Users compose very short tweets akin to SMS messages [Battestini et al., 2010]. This takes shape mainly in the form of abbreviated '*text-speak*' e.g. "`how r u LOL`".

2. **This spike on the tail end, at 140 characters:** The spike at 140 characters is due to the truncation of messages by bots and other Twitter content-generating programs; and regular users trying to maximize the use of all 140 characters while composing a tweet. Abbreviations are employed to fit the tweet in the character limit.

With my results from Experiment 5.12, and also prior literature [Yoshida et al., 2010; Zarrella, 2009] in mind, I conclude that the characteristic bimodal distribution of tweet lengths is indicative of everyday, real-world chatter, representative of the *10-Gigabyte Dataset*.

5.6.2 Message Entities — Replies, Retweets, Hashtags, and URLs

In the study of summary statistics with respect to tweet content, I have also surveyed the frequency of occurrences of entities (defined in Sections 4.8.2 and 4.8.3) in Twitter messages. Experiment 5.13 documents the distribution of such entities in the real-world *10-Gigabyte Dataset*, which includes: `@user` notations, `RT`s (retweets), `#hashtags`, and URLs.

Experiment 5.13. *Large-scale investigation on the distribution of entities within tweet content in the 10-Gigabyte Dataset.*

METHOD: Using the string-extraction methodology introduced in Sections 4.8.2 and 4.8.3, entities found in each message in the *10-Gigabyte Dataset* — `@user` notations, `RT`s (retweets), `#hashtags`, and URLs — are extracted and counted. A frequency distribution is obtained, from which the overall tweeting habits from the Twitter user base (representative of the *10-Gigabyte Dataset*) can be inferred, based on the ideas proposed in current literature on Twitter messaging.

Results and Discussion: Figure 5.15 illustrates the distributions of `@user` notations, RTs, `#hashtags`, and URLs, respectively, in Twitter message content.

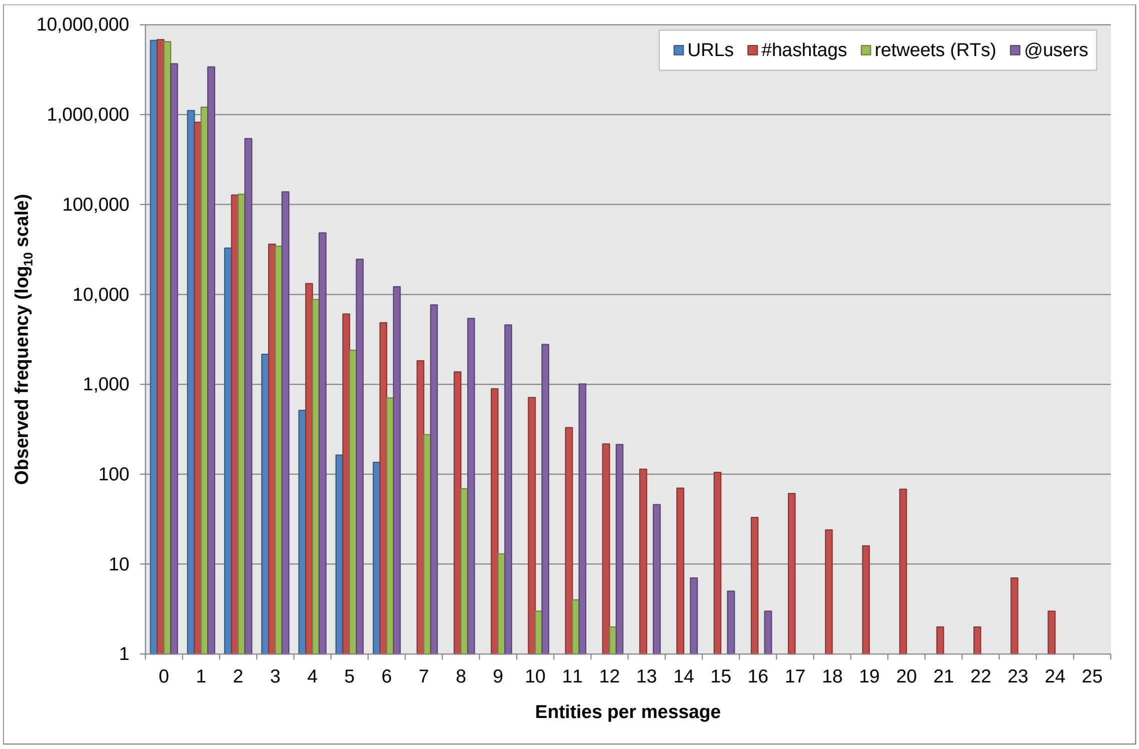


Figure 5.15: Histogram of per-message occurrences of `@user` notations, as a logarithmic plot, of the *10-Gigabyte Dataset*.

From the distributions (Figure 5.15), two points of discussion could be raised:

- **Outliers:** Outliers were present in the frequency distribution of every entity, save for URLs. Twitter users, especially newly-registered ones, sometimes misuse entities in a tweet as they do not know the specific format or purpose of such entities. Examples were found in the distributions of `@user` notations (one tweet with 19 `@user` entities was found) and `RT` notations (one tweet with 20 retweet entities was found). As for `#hashtags`, outliers in the distribution are caused by spam tweets, where the presence of many tags in a single tweet creates a higher likelihood of message visibility; seven outliers containing 28 or more `#hashtags` in a single message were removed.

- **Distribution:** Every one of the four entities conforms to a power-law distribution. When modeled with a power-law distribution, each entity exhibits a high goodness-of-fit, with Pearson R^2 values (Equation 5.2) of the entities `@user`, `RTs` (retweets), `#hashtags`, and URLs amounting to $\{0.9622, 0.9418, 0.8461, 0.9599\}$ respectively.

Co-occurrences of Entities in Messages

I augment the results of Experiment 5.13 with Experiment 5.14, which documents the co-occurrences of entities in tweets. The study of entity co-occurrences will provide a big-picture on Twitter *messaging intents* and *messaging genres*.

Experiment 5.14. *Large-scale investigation on the co-occurrences of multiple entities within tweets in the 10-Gigabyte Dataset, to identify the messaging intents and messaging genres in real-world Twitter usage.*

Method: For this experiment, the string-extraction methodology introduced in Sections 4.8.2 and 4.8.3 is again applied to extract entities from each tweet in the *10-Gigabyte Dataset*.

Instead of counting the number of entities per tweet (Experiment 5.13), I will count the co-occurrences of the four types of entities. This will result in a total of 16 (= 4^2) possible combinations of entity co-occurrences. To illustrate: a tweet, say, "`@bob Are you getting #pizza?`" has an entity co-occurrence of `@user`+`#hashtag`.

Results and Discussion: Table 5.6.2 lists all possible co-occurrences of entities in Twitter messages, and the relative proportion of such messages within the dataset.

Description	**Count**	**Percent**	`@user`	`#hashtag`	`RT`	**URL**
Normal text	2,540,213	34.42%				
URL only	601,488	8.15%				✓
`RT` only*	2,967	0.04%			✓	
`RT` and URL*	1,305	0.02%			✓	✓
`#hashtag` only	397,348	5.38%		✓		
`#hashtag` and URL	133,733	1.81%		✓		✓
`#hashtag` and `RT`*	1,290	0.02%		✓	✓	
`#hashtag`, `RT` and URL*	815	0.01%		✓	✓	✓
`@user` only	2,431,653	32.95%	✓			
`@user` and URL	155,925	2.11%	✓			✓
`RT` and `@user`	942,198	12.77%	✓		✓	
`RT`, `@user` and URL	171,228	2.32%	✓		✓	✓
`@user` and `#hashtag`	182,065	2.47%	✓	✓		
`@user`, URL and `#hashtag`	32,205	0.44%	✓	✓		✓
`RT`, `@user` and `#hashtag`	223,741	3.03%	✓	✓	✓	
All four	45,476	0.62%	✓	✓	✓	✓

Table 5.6: Table illustrating the 12 possible co-occurrences of message entities — `@user` notations, `RT`s, `#hashtags`, and URLs — and their relative frequency among messages in the *10-Gigabyte Dataset*. Rows denoted with an asterisk (*) are combinations which are not following Twitter messaging convention, as the retweet notation must refer to a username (i.e. "`RT @user`").

Several conclusions can be drawn from the data in Table 5.6.2, based on existing findings in extant literature on tweeting habits [Huang et al., 2010; Honeycutt and Herring, 2009; Boyd et al., 2010; Hughes and Palen, 2009; Starbird et al., 2010; Longueville et al., 2009; Mendoza et al., 2010], as earlier mentioned in the literature review (Chapter 3).

Also, drawing from the theory of *messaging genres* [Westman and Freund, 2010; Battestini et al., 2010; Ehrlich and Shami, 2010; Ritter et al., 2010; Sriram et al., 2010; Subramanian and March, 2010], I am also able to explain the motivations behind specific entity co-occurrences within tweets from a user's perspective.

- Firstly, a total of ~0.1% of the tweets that I studied in the *10-Gigabyte Dataset* contain combinations of entities which does not follow Twitter convention. These co-occurrences of entities, indicated by an asterisk in Table 5.6.2, contain retweet notations which do not include a username (i.e. "`RT @user`"), and is illustrative of users **not knowing how to use Twitter properly**.

- ~34.42% of tweets contain no entities whatsoever: these belong to the **personal updates** genre of messages, involving sharing of personal information, personal opinion, and daily chatter or a 'status update' to their close followers [Westman and Freund, 2010; Subramanian and March, 2010; Ritter et al., 2010; Sriram et al., 2010].

- ~32.95% of tweets contain only `@user` entities, signifying Twitter usage for **directed dialogue**: conversations/dialogues/questions, addressed to certain user(s), and mostly part of a thread of conversation [Westman and Freund, 2010; Battestini et al., 2010; Ritter et al., 2010; Ehrlich and Shami, 2010].

- ~18.12 % of tweets include retweets (combined with `@user` notation as per convention), which may include `#hashtags` and URLs. This demonstrates the **real-time sharing** of tweets [Westman and Freund, 2010], where it is commonly reflected in users rebroadcasting popular posts, news, and links via retweets [Ritter et al., 2010; Ehrlich and Shami, 2010]

- ~12.51% of tweets contain URLs (with combinations of other entities) which are *not* part of retweets. This refers to the sending of URLs to another user, or broadcasting a URLs to other people in general. This is commonly used in **business broadcasting** [Westman and Freund, 2010] for product deals and brand promotion [Westman and Freund, 2010; Subramanian and March, 2010; Sriram et al., 2010]. A section of tweets in this genre include spam/marketing behavior [Horn, 2010] based on their study of Twitter accounts by 'companies' with such characteristics. There are exceptions to this broad genre, such as the publication of bona fide tweets containing URLs to useful websites or social media.

- ~7.85% of tweets are **personal updates** or **directed dialogue** tweets (as above), albeit with the presence of `#hashtags`. In such cases, of `#hashtags` only serve to augment their original message, in terms of grouping conversations in context [Longueville et al., 2009], improve the discovery of related messages [Golovchinsky and Efron, 2010], and participation in Twitter memes [Huang et al., 2010].

5.7 Concluding Notes

Before I proceed to the next chapter dealing with the clustering of Twitter metadata and inferences, I will recap what has been covered in this chapter, in terms of my contribution to large-scale explorations of Twitter.

Segueing from Chapter 4, the initial part of this chapter covered two frameworks that I developed for the harvesting and extraction of Twitter metadata. *Twitmographics* was first designed for small-scale data gathering of archived/backdated Twitter data using the on-demand REST-`user` API. With the maturation of the Streaming API for large-scale real-time data collection, *TweetHarvester* was developed to meet the need for a data-gathering framework that can successfully capture metadata in the order of millions.

Naturally, following from my introduction of the two frameworks, I then explained how such frameworks can be used to obtain a large-scale representative dataset — comprising of real-world tweets, users, and metadata — with emphasis on the large-scale availability and ease of acquisition of such data using the Streaming API. Thus, the *10-Gigabyte Dataset* was created in 2011, which was to be used in large-scale experiments and exploration of the properties of the entire *Twitterverse*.

Armed with the *10-Gigabyte Dataset*, and the knowledge of how raw metadata can be distilled to meaningful information, the latter part of this chapter was devoted to applying those metrics, algorithms, and analyses proposed earlier in Chapter 4 onto the *10-Gigabyte Dataset*. Analyses on the real-world *10-Gigabyte Dataset* have been conducted from the perspective of real-life demographic properties: gender, location, device class (Section 4.6); online presence: profile customization, web presence, user connectivity, and loyalty/usage frequency (Section: 4.7); and communication patterns: message length analysis, and message entity analysis (Section 4.8). Such analyses have provided an insight of the Twitter user base and their tweets, in terms of online latent behavior and also real-world characteristics.

Moving ahead from the current chapter, Chapter 6 will investigate how both the raw metadata and the inferences proposed in Chapter 4 — and demonstrated on large real-world data in the current chapter — could be used together to reveal hidden traits in certain segments of the user base. The suitability of multiple data mining techniques will be examined with respect to the clustering of microblogging data. Also in the following chapter, the outcomes from combining pattern recognition with my Twitter knowledge-discovery approaches will be described in detail.

Chapter 6

Clustering Twitter Metadata and Inferences

> *"You say that we've got nothing in common*
> *No common ground to start from and we're falling apart...*
> *And I said, 'What about Breakfast at Tiffany's?'*
> *She said, 'I think I remember the film, and...*
> *as I recall, I think we both kinda liked it'*
> *And I said, 'Well, that's the one thing we've got'*
>
> — Deep Blue Something
> *Breakfast at Tiffany's* (1995)

Parts of this chapter have been published as:

Cheong, M. and Lee, V. [2009]. Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base, Proc. CIKM 2009 Co-Located Work-shops: SWSM 2009, pp. 1–8.

Cheong, M. and Lee, V. [2010b]. A Study on Detecting Patterns in Twitter Intra-topic User and Message Clustering, *Proc. ICPR 2010*, pp. 3125–3128.

Cheong, M. and Lee, V. [2010c]. Twitmographics: Learning the Emergent Properties of the Twitter Community, *From Sociology to Computing in Social Networks: Theory, Foundations and Applications, Vol. 1 of Lecture Notes in Social Networks*, Springer-Verlag, pp. 323–342.

Cheong, M. and Lee, V. [2011]. A Microblogging-based Approach to Terrorism Informatics: Exploration and Chronicling Civilian Sentiment and Response to Terrorism Events via Twitter, Information Systems Frontiers **13**(1): 45–59.

In the previous chapters, I have discussed how Twitter's user and message metadata are rich sources of data to be mined. From such data, inferences on users' online behavior, their social characteristics, and real-world demographics can be revealed using algorithms and metrics. I have also designed prototypes for the collection of such metadata, and successfully gathered a large-scale dataset from the real world. My proposed inference methods were applied on real-world data to learn about the latent properties hidden within real-world Twitter metadata.

In this chapter, I will study how simple but useful patterns can be identified among both the raw metadata and resulting inferences. I also apply existing pattern recognition algorithms to reveal patterns which aren't readily obvious at first glance, nor attainable by observing individual inferences.

I will firstly introduce the concept of clustering using Kohonen's Self-Organizing Map (SOM) algorithm (Section 6.1). Within this section, I will detail related research on SOM-based clustering on social network data, which is relevant to Twitter as a form of online social network. I will also detail the *Viscovery SOMine* data mining and knowledge discovery package which couples the SOM algorithm with Ward's clustering algorithm for fast and effective clustering, visualization, and cluster analysis. The rationale for choosing this data mining package will also be discussed.

In Section 6.2, I will detail an original study done in the early stages of my research that deals with SOM clustering on manually-extracted metadata and inferences on Twitter Trending Topics (introduced in Section 4.1.4); in which I will describe the dataset used, the resulting clusters, and discuss how the clustering results relate to the nature of the topics.

Following that, I shall conduct a similar empirical study in Section 6.3. This time, the study will be based on a dataset harvested automatically using the *Twitmographics* prototype 5.1, followed by the automated generation of inferences from raw metadata.

In the final section (Section 6.4), for the sake of completeness, I will compare SOM-Ward-based clustering with k-means (another widely used approach to clustering); the differences between the approaches; as well as the evaluation of results from both approaches. To conclude this chapter, I will provide a discussion on the suitability of SOM-Ward clustering specifically for Twitter inferences.

6.1 Self-Organizing Maps for Inference Clustering

6.1.1 Primer on Clustering

Clustering is "the organization of a collection of patterns into clusters based on similarity" [Abbas, 2008]. The analysis of clusters is "an important technique [in] ...exploratory data analysis [Jain and Dubes, 1988]. Data is organized "by abstracting underlying structure" in a representation that can "be investigated to see if the data group [is in accordance with] ...preconceived ideas or to suggest new experiments" [Jain and Dubes, 1988]. Also, one could explore the "structure of the data that does not require the assumptions common to most statistical methods" [Jain and Dubes, 1988]. The applications of clustering and cluster analysis have been covered in several popular texts cf. [Hartigan, 1975; Anderberg, 1973].

6.1.2 About Self-Organizing Maps

In my first paper [Cheong and Lee, 2009], I proposed the usage of Self-Organizing Maps (SOMs) for clustering of Twitter metadata, and inferences generated from said metadata. To my knowledge to date, there was no precedence of applying such a technique on Twitter-based inferences. However, SOM-based clustering has been applied on other forms of social network data (Section 6.1.3).

The SOM, first introduced by Kohonen [1988], is an unsupervised visual clustering technique which works by projecting input from higher dimensions into maps of two-dimensions. In the projected 2D maps, similar features are spatially close by on the map, which is an effective method of clustering and visualizing clustered data [Mitra and Acharya, 2003].

For brevity, I will present the basic principles of SOM clustering as first proposed by Kohonen [1988] in algorithmic form (Algorithm 6.9).

Algorithm 6.9 An algorithmic overview of the Kohonen [1988] Self-Organizing Map algorithm

```
procedure KohonenAlgorithm
Require: a set of input vectors $x \in \mathbb{R}^n$, where $n$ = number of dimensions.
Require: a set of $i$ nodes in map, each with model vector $m_i(t) \in \mathbb{R}^n$, $t$ = current iteration.
Require: learning-rate factor $\alpha(t)$, decreasing monotonically with the regression steps.
Require: neighborhood function $h_{ci}(t)$, that determines radius around a given node in the SOM.
Require: $\lambda$, an upper limit on $t$
    $t \leftarrow 1$                                        ▷ $t$ = the current iteration.
    Initialize $m_i(1)$                                     ▷ Initialization of the map.
    while $t < \lambda$ do
        pick an input vector $x$
        for all model vectors $\vec{m_i}$ do
            if $|m_i - x| \leq |m_c - x|$ then
                $m_c \leftarrow m_i$                        ▷ $c$ is the node with best match to $x$.
            end if
        end for
        $m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_i(t)]$    ▷ Update nodes around $m_c$.
        $t \leftarrow t + 1$
    end while
end procedure
```

In the case of Twitter metadata (in both user and message domains), similar users are grouped together in the resulting SOM. For instance, in a given neighborhood in the SOM, users will share similarities such as in messaging style/behavior, geographic location, and demographic properties.

6.1.3 Rationale for using SOMs

Prior work on using SOMs to classify and visualize data from social networks are commonplace. Several examples include the following, where the latter two are especially relevant to my work on Twitter metadata and real-world demographic inferences:

- **Clustering MySpace music artists**: SOMs were used in clustering music artist profiles on MySpace to represent their popularity patterns of Myspace based on "their attributes on the platform and their position in the social network." [Couronné et al., 2009]. Further work in the same vein as [Couronné et al., 2009], Stoica et al. [2010] used SOMs to cluster artists on a social network based on their "audience and authority characteristics." They explicitly make use of MySpace user metadata — not unlike my use of Twitter metadata in the user domain — which comes in the form of MySpace profile visits, comments, 'best friend' status on MySpace, friendship mutuality, and artist categorization.

- **Visualization and grouping of Habbo Hotel users**: Clustering of the online Habbo Hotel chat community/social network, in order to visualize groups and the distribution of 'buddies' related to a user [Torres, 2004], has been done using a SOM. Torres [2004] was able to cluster over 50,000 users; the SOM was particularly advantageous over other methods as this method of "...visualizing data is extremely efficient at showing a very large amount of information in a relatively small space" [Torres, 2004].

- **Clustering blog communities**: Merelo-Guervs et al. [2004] uses a SOM to map the indegree and outdegree statistics of blogs (edges represented by links to other blogs) — not unlike my use of the follower/friend metric in the Twitter user domain — to cluster blog 'communities'. Similar work was done by Jamali and Abolhassani [2007], albeit by using the outdegree correlation vector of blog links.

- **Cluster analysis of sociodemographic properties from *mixi* data**: Using data from the Japanese social network *mixi*, [Yamazaki and Kumasaka, 2010] obtained data from online communities in terms of place/location names, regional dialects, food, sports, schools, and other properties. Using SOMs, they have identified five distinct clusters of prefectures, based on the characteristics of their inhabitants.

- **Analysis of tourist-related tweets in the message domain**: Claster et al. [2010] performed experiments on a real-world Twitter dataset of 70 million tweets to cluster tourism-related tweets. By extracting tourism-related keywords from messages after sanitization (based on their own lexicon), they aggregate such tweets into groups, based on sentiment found in tweets, using a hybrid algorithm based on an artifical neural network, combined with a SOM for "to provide a richer understanding of the data" [Claster et al., 2010].

6.1.4 The *Viscovery SOMine/Profiler* Package

For visual clustering of Twitter data — both raw metadata and derived data from my inference algorithms — I utilized the *Viscovery SOMine* software by *Eudaptics GmbH* (which is also available as a component of the *Viscovery Profiler* data mining package).

I used *Viscovery SOMine* to perform experiments on SOM clustering, as prior research found that visualization and analysis of data can be done "...without any prior statistical knowledge of the data set" [Li, 2005]. Furthermore, before constructing the SOM, *Viscovery SOMine* also "provides suggestions as to which data items should be grouped together" [Li, 2005]; and it simplifies the process of fine-tuning the results and controlling data processing [Li, 2005].

Also, *Viscovery SOMine* has been found to effectively cluster large volumes of demographic data: of the order of ~1.5 million records consisting of 10 dimensions each [Yao et al., 2010]. According to the developers, *Viscovery SOMine* can potentially classify up to "...50,000 previously unseen data records" per second [Eudaptics Software GmbH, 2005]. It is able to achieve a high speed of SOM clustering due to the usage of an optimized version of Kohonen's batch SOM algorithm [Yao et al., 2010; Eudaptics Software GmbH, 2005]; map growing optimizations, and usage of heuristics (a gradient-descent algorithm) to approximate a best-match during SOM construction [Eudaptics Software GmbH, 2005][1]

For visualization and analysis of the cluster distributions, *Viscovery SOMine* provides three forms of visualizations: SOM-Ward, Ward, and SOM Single Linkage [Yao et al., 2010; Deboeck and Kohonen, 1998]. Among the three, SOM-Ward is chosen due to its effectiveness [Yao et al., 2010] over other methods. Using *Viscovery SOMine*'s implementation of SOM-Ward, the input dataset is "...first projected onto a two-dimensional display using the SOM [local-ordering], and the resulting SOM is then clustered" [Yao et al., 2010] with Ward's agglomerative hierarchical clustering (i.e. 'bottom-up') method [Yao et al., 2010; Ward, 1963]. Ward's approach [Ward, 1963] "outperforms other hierarchical clustering methods" with respect to squared-error differences [Jain and Dubes, 1988].

In brief, SOM-Ward starts with a clustering where "each node is treated as a separate cluster" [Yao et al., 2010]. In subsequent iterations, two clusters with the minimum Euclidean distance — "...[taking] into account not only the Ward distance but also the topological characteristics of the SOM" [Yao et al., 2010] — are merged in each step. Distance between two "...non-adjacent clusters is considered infinite" [Yao et al., 2010], resulting in the merging of adjacent clusters. SOM-Ward offers flexibility to the researcher as the selection of "the most appropriate number of clusters" [Yao et al., 2010] can be fine-tuned.

6.2 Initial Study on SOM Suitability for Twitter Metadata

SOM-based clustering can provide one with an idea of how users (and their authored messages) can be categorized and clustered based on certain distinctions evident in their

[1] Technical information on the inner workings and optimizations employed in *Viscovery SOMine* is available in [Eudaptics Software GmbH, 2005], or at <http://www.viscovery.net/faqs>.

metadata, and inferences regarding said metadata. In other words, the clustered data can provide representative characteristics of the users contributing to a particular Twitter topic.

In my initial published paper [Cheong and Lee, 2009], I experimented with the idea of clustering to find meaningful patterns in Twitter data from a small dataset harvested using the old REST-`user` and Search APIs (Section 5.1), based on certain topic keywords. For brevity, the dataset resulting from the study in this section will hereinafter be referred to as the *SWSM2009 Clustering Dataset*, referring to the conference in which the results were published in.

6.2.1 Preliminaries

This preliminary study [Cheong and Lee, 2009] involved selecting two topic keywords for different types of messages (based on popularity and distribution over time), bearing in mind the limitations of the `search` API in terms of data collection.

Each of the cases belong to one of three categories [Cheong and Lee, 2009]; a full description of each category is in Section 8.2.1:

1. **Long-term topics:** Sparsely-discussed topics, due to their relative obscurity.

2. **Medium-term topics:** Topics which are either: generic terms which are common but do not warrant a high number of tweets, or those exhibiting sustained ‘trailing patterns’, i.e. part of a long-term discussion that is slowly decreasing in momentum [Fukuhara et al., 2005].

3. **Short-term topics:** Such topics are high-volume in nature and can be a very commonly-talked about term which does not exhibit spiking behavior; or also be high-volume topics captured using the Twitter API in the middle of a spike.

A treatise on studying the trending behavior of messages in each category above is provided in Section 8.2. For now, it suffices to analyze the properties and features of the tweets themselves as opposed to studying their temporal trending behavior.

Table 6.1 details the distribution of keywords/topics within the *SWSM2009 Clustering Dataset*.

One thing noticeable in Table 6.1 is the small number of samples in each category (and collectively, the entire dataset). This is due to two reasons; firstly, the fact that the Search and REST-`user` API is severely restrictive in terms of volume of results returned (Section 4.1.3). Secondly, as will be elaborated in the next subsection, my initial study on user inferences and demographic information on Twitter was done manually without any automated algorithms. As such, sampling was performed to simplify the process of manual coding and information extraction [Cheong and Lee, 2009].

6.2.2 Experimental Analysis

The methods and results of the clustering exercise conducted in [Cheong and Lee, 2009] is documented in Experiment 6.1.

Table 6.1: Distribution of topics/keywords in the *SWSM2009 Clustering Dataset*.

Keyword	**Type**	**No. of messages**	**Description**
`Grey's Anatomy`	Trending (15 May 2009), short-term topic.	109	This refers to a television drama series which just had its season finale aired on primetime television.
`H1N1`	Trending (first peaked on 1–2 May, again on 15 May), medium-term topic.	101	H1N1 refers to the Swine Flu pandemic.
`Nizar`	Trending (11 May 2009), long-term topic.	61	Nizar is the name of a politician involved in a constitutional crisis in a Malaysian state [Mageswari and Goh, 2009].
`TwitHit`	Trending (15 May 2009), medium-term topic.	81	This keyword appears as a result of Twitter users who supply their login credentials to a dubious spamming site [Cashmore, 2009b].
`Coffee`	Non-trending (control) topic.	111	Common everyday topic, included here for comparison.
`Revolver-held`	Non-trending (control) topic.	21	Relatively obscure topic (name of a German alternative rock band); included here for comparison.
Total		484	

Experiment 6.1. *SOM-Ward clustering with Viscovery SOMine to find meaningful patterns in Twitter data from the SWSM2009 Clustering Dataset.*

Method: In the *SWSM2009 Clustering Dataset*, I randomly sampled approximately 13% of the total tweets from unique users, resulting in a total of 484 records, as distributed across the 6 topics in Table 6.1. Sampling was performed as the inferences from raw metadata in this early experiment were obtained through the manual browsing and observations of Twitter user profiles, without any automation whatsoever [Cheong and Lee, 2009].

Useful information is inferred from the tweets' authors by visiting the authors' Twitter profile page. Note that the data are not verified against any third-party source but taken as-is written by the users themselves [Cheong and Lee, 2009].

A total of four features were selected and manually coded [Cheong and Lee, 2009], viz.:

1. **Client and device used** is available directly from the tweets obtained. The device (computer, mobile phone, or culled from external data sources) can be ascertained from the codename of the Twitter client application used (a similar method was applied by Java et al. [2009] with respect to this). (Section 4.6.3 details my algorithm that evolved from this concept).

2. **Gender**, in this experiment, is identified by the writing style of the user (e.g. "`@username misses her friends`" indicating the female gender). If such cues are

absent, inspection of the profile information (primarily first name), and also the profile image that is publicly available on a Twitter profile is used to determine gender. Besides the male and female sexes, a third category, the *neuter* gender is included for Twitter users that are created by a workgroup or an organization. (Section 4.6.1 details my gender-inference algorithm that automates this process from first name information in user metadata).

3. **Primary usage pattern** is deduced by manually reading through the first page of a users tweets. If this was inconclusive, I manually conducted a visit to the users homepage, publicly available as a link on their profile. The usage patterns fall into one of five categories: *personal* (majority of the postings are personal communication and social networking; examples would be messaging friends, sharing information); *group* (a not-for-profit user group with common interest, such as fan clubs or groups in which researchers network); *aggregator* (publishing or collating information as part of their job — for example news agencies, Twitter accounts linked to RSS feeds, politicians message to their constituents — with little or no personalized content or messaging performed; *satire* (Twitter account for humorous, satirical, or parody purposes); and lastly *marketing* (Twitter account to 'push' a product; the majority in this category comprise of spam, unsolicited postings, and possibly harmful sites). I have since introduced inference algorithms for determining device classes and deducing usage behavior vis profile URL metadata, discussed prior in detail in Section 4.6.3 and 4.7.2 respectively.

4. **Country** is based on the user profile's *location* field (which can take a form of a city such as '*Adelaide*'; or a GPS-generated coordinate pair. Sometimes, the 'location' is deduced from the country code found in the user's profile URL. This process has since been automated using my algorithm as detailed in Section 4.6.2.

SOM Clustering Parameters

The usage pattern of Twitter among users, in each of the six topics listed in Table 6.1 above, have been fed into *Viscovery SOMine*. All the sample data were used as the training data for *SOMine*, while applying the program's default parameters (Table 6.2) for generating SOMs for clustering.

For each topic covered in Table 6.1, a *preliminary model* (comprising of a map and "an optimal SOM-Ward clustering that is allocated in a segmentation" [Eudaptics Software GmbH, 2005]) is created. For visualization, individual maps for each feature vector that can be generated; these maps provide a graphical representation of how each feature is represented within a particular cluster.

For example, Figure 6.1 illustrates three maps for discrete values of the primary usage pattern feature — *personal*, *aggregator*, and *marketing* —for the Twitter topic `coffee`. Observe how the data points corresponding to different *primary usage patterns* are distributed within the two clusters (which are demarcated by a black border).

Table 6.2: Initial parameters for SOM clustering in *Viscovery SOMine*. These parameters are "well-defined, proven default settings" determined by the publisher [Eudaptics Software GmbH, 2005] which have worked well in prior research e.g. [Li, 2005].

Parameter	**Value**	**Description**
Map Size	1,000	Number of nodes in the map, determines granularity.
Map Format	Automatic	Determines the aspect ratio of the resulting map; the *Automatic* option derives it from the "...ratio of the principal plane of the source data set." [Eudaptics Software GmbH, 2005]
Tension	0.5	Smaller tension values mean greater adaptation of the SOM to the data space ("the more the differences in the data are represented and the less the attribute values are averaged") [Eudaptics Software GmbH, 2005].
Training Schedule	Normal	Internally used by *Viscovery SOMine* to determine the speed, number of iterations and accuracy of the results; this ranges from *Fast* (least accurate, quickest), *Normal*, to *Accurate* (most accurate, slowest).

Banned or suspended accounts were included in the following cases as a separate discrete category, with its features substituted with an X. Information which cannot be deduced is denoted with a *.

Results and Discussion: A discussion of the obtained clusters, with qualitative evaluation of the obtained clusters are as follows, organized by the type of topic.

Long-term topics

The SOM for control (non-trending) topic Revolverheld reveals that the majority of the users contributing to the chatter (blue cluster) are females in Germany who mainly contribute personal chatter on Twitter using the web interface.

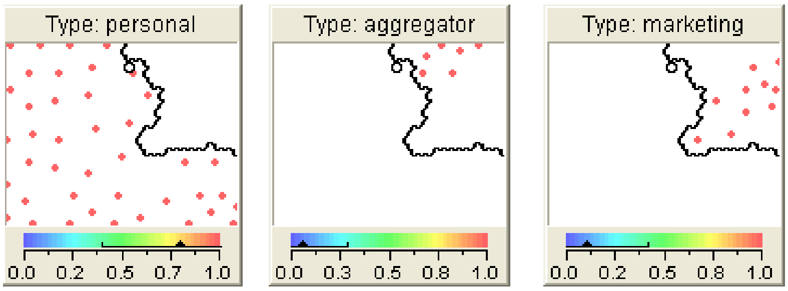


Figure 6.1: An example of three maps (for the coffee topic), representing discrete values of *primary usage pattern*. The overall clustering consists of two main clusters, separated by a black border in terms of the 2D map space. From left to right: *personal*, *aggregator*, and *marketing*. Note the distribution of the values within each cluster.

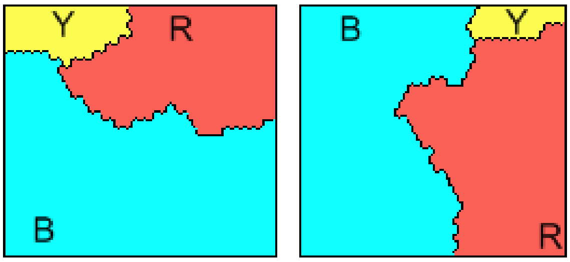


Figure 6.2: SOM clusters generated for the long-term topics (a, left) `Revolverheld` and (b, right) `Nizar`.

The red cluster represents German males/organizations which aggregate news regarding the `Revolverheld` band using social media clients; and the yellow cluster depicts anonymous users (with no geographic location nor gender information accessible) contributing to the discussion anonymously.

For the long-term trending topic `Nizar`, the majority of the conversation is generated by Malaysian users (relevant, since the topic is a Malaysian news story) of both genders who mainly use Twitter for personal microblogging. The red cluster consists of predominantly males from other countries, using Twitter as a form of citizen journalism to aggregate and publish news. It is interesting to note that there are a proportion of users (the smallest cluster) which are organizations which either aggregate data or perform aggressive marketing while piggybacking on a Trending search term; almost all feeds from this category of users are culled from RSS feeds.

Medium-term topics

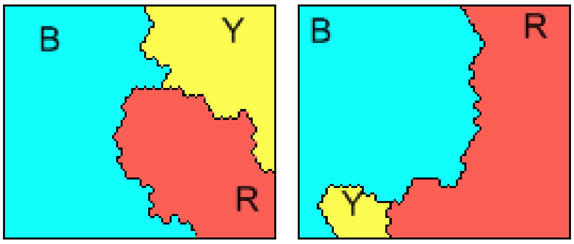


Figure 6.3: SOM clusters generated for the medium-term topics (a, left) `H1N1` and (b, right) `TwitHit`.

For the medium-term `H1N1` trending topic (which is a global affair), several interesting trends can be observed from the generated Kohonen SOM. The blue cluster comprising the majority of the user sample comprise of male Twitter web users who microblog about personal matters, situated in Malaysia, the United States, and other countries in Asia genuinely tweeting about the flu pandemic. The yellow cluster consists predominantly of news aggregators (by organization-based Twitter accounts) sourcing data from RSS feeds that contribute to the heavy hype about the flu pandemic on Twitter; what brings attention to this cluster is that a subsection of this consists of users whose accounts have been banned by Twitter for account violation. The red cluster is closely related to the previous cluster, where the majority of users singled out in this cluster are marketing-based anonymous Twitter users including banned accounts whose sole modus operandi is piggybacking on the `H1N1` topic for spamming and deceitful advertising purposes.

Studying the SOM for `TwitHit` reveals the demographics behind users who fall prey to internet scamming/spam-based sites. The red cluster represents the majority of users falling prey to the scam — majority of users in this cluster are American regular Twitter users of both sexes who use Twitter for typical personal microblogging. The blue cluster represents dubious accounts which have the US and the UK as country of origin, using Twitter for mostly aggressive marketing and spamming activities — which possibly indicates the root cause of the problem. A small cluster of Australian-based users of Twitter who use Twitter as a form of social networking and also personal microblogging are the next affected set of users outside of countries on the Western side of the globe.

Short-term topics

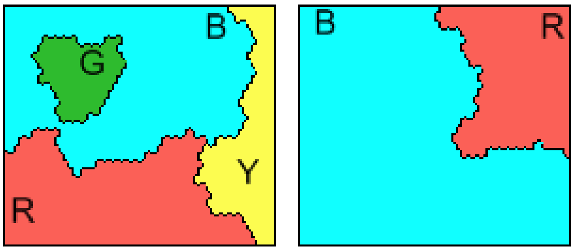


Figure 6.4: SOM clusters generated for the short-term topics (a, left) `Grey's Anatomy` and (b, right) `coffee`.

I will now analyze the results for the short-term trending topic `Grey's Anatomy` (for the US drama series). The cluster in red successfully reflects the demographics of the drama series — female Twitter users based in the United States, using Twitter mainly for personal microblogging. However, this cluster reveals some of the demographics that are not obvious from cursory inspection — the majority of the Twitter users in this cluster actively contribute to Twitter not only through the web, but rather mostly through other social media clients and also using mobile clients, indicating a shift from traditional usage of Web 2.0 services such as Twitter from the desktop web environment to a more mobile, social-based environment [Boyd and Ellison, 2007].

The blue cluster represents users (also predominantly female and use Twitter for personal communication), however their geographic location spans the continents of Asia and Europe and their main contribution to the microblog chatter comes mainly from the web interface. The remainder of chatter on the drama series from the US — marked as a yellow cluster - comes from aggregator users, that either exhibit characteristics of collating news feeds from the entertainment/television industry, or part of marketing spammers piggybacking on a trending term. The final cluster (in green) also reveals demographic data that isn't readily apparent; this group consists of (predominantly) female Twitter users using Twitter as a personal communications medium, entirely from Canada who use a hybrid of methods of posting to Twitter. Data such as the ones illustrated above are highly valuable to people in the media, advertising, and television production industry, which illustrates the motivation behind this research.

As for the `coffee` keyword, the majority of users are from the UK and the US from both sexes, who tweet about coffee in a personal context, describing part of their daily routine. The *SWSM2009 Clustering Dataset* was indeed collected during breakfast time in the GMT+0 time zone, justifying this phenomenon. The remainder of the sample set (in red) comprises of Twitter user accounts involved in coffee-related marketing campaigns and news aggregation; some of the accounts in this cluster have been suspended or banned based on policy violation.

6.2.3 Study Conclusion

From the examples shown in these preliminary studies [Cheong and Lee, 2009], SOM-based clustering allows visual categorization and clustering of users contributing to a trend based on their demographic data. This could potentially be useful in decision-support (e.g. policy-making and socio-economic planning), where the clustered data can provide representative characteristics of the users contributing to a particular Twitter topic. Whilst the detection of hidden patterns for topics of interest is also possible based on qualitative and subjective judgement, experimental results reveal that clustering provides meaningful interpretations of certain clusters of Twitter metadata.

The next section documents an analysis of SOM-Ward clustering on a larger dataset created using automated approaches, contrasting with Experiment 6.1 in terms of size and inference method.

6.3 Combining SOMs with *Twitmographics*' Automated Inferences

The previous section (Section 6.2) dealt with the clustering of the *SWSM2009 Clustering Dataset* dataset, which is a small-scale dataset containing manually-obtained inferences and statistics. Hence, in the current section, my focus is to study if SOM-Ward clustering is equally as effective, if not better, when applied to a larger dataset, obtained using automated approaches to generate inferences (Sections 4.6, 4.7, and 4.8).

6.3.1 Experimental Goals and Design

In Section 5.1, I have designed the *Twitmographics* prototype for automated harvesting of user and message metadata based on the old Twitter REST and Search APIs (Section 4.1). This section details an investigation on feeding the inferences and summary statistics (fully discussed prior in Chapter 4) automatically generated from my *Twitmographics* prototype into *Viscovery SOMine* for clustering. This was initially published in [Cheong and Lee, 2010c], and analyses on the case studies further described in [Cheong and Lee, 2010b].

I tested the exploratory framework on several global-and regional-concern trending topics, namely:

Table 6.3: Topics used and overview statistics in the *Twitmographics* study [Cheong and Lee, 2010c,b].

Search term	**Total messages (excluding bans)**	**Banned users**	**Unique users (excluding bans)**
`Iran Election`	4,905	0	1953
`iPhone`	4,246	2	3,368
`Obama`	4,640	5	3,115

1. The 2009 Iran Election issue [Fleishman, 2009] (keyphrase: `Iran Election`)

2. US President Obama's reaction toward the Iran issue and foreign policy (keyword: `Obama`)

3. The iPhone OS 3.0 software launch [Martin, 2009] (keyword: `iPhone`)

The rationale behind the topic selections above is to observe the pattern of Twitter interaction by its users with regards to current affairs, political, and technology topics. The *Twitmographics* prototype-generated data is again visualized — as per Cheong and Lee [2009], also discussed in the prior Section 6.2 — using the SOM-Ward-Clusters approach in *Viscovery SOMine*.

Again, the objective is to reveal the emergent properties behind the Twitter user base expressing their views on the abovementioned topics [Cheong and Lee, 2010c], this time with emphasis on higher dimensionality of features (more metadata was analyzed as the study on metadata inferences developed), and with a larger dataset (which was hitherto impractical due to the amount of manual analyses performed, cf. [Cheong and Lee, 2009]).

Table 6.3 summarizes the keywords, topics, and vital statistics of the accumulated corpus of data, which will be named the *Twitmographics Clustering Dataset* for brevity.

6.3.2 Experimental Analysis

The SOM-Ward clustering experiment using *Viscovery SOMine*, together with its results, are documented in Experiment 6.2.

Experiment 6.2. *SOM-Ward clustering with Viscovery SOMine to find meaningful patterns in Twitter data and evaluate the efficiency of clustering on automatically-generated inferences, on the Twitmographics Clustering Dataset.*

Method: As per the initial study [Cheong and Lee, 2009] in Section 6.2, all sample data were used as the training data. *Viscovery SOMine*'s default parameters (again, cf. Table 6.2) were used in the SOM clustering process.

Results and Discussion: The results from the clustering are as follows:

Case 1: `Iran Election`

Figure 6.5 shows the high-level results of automated clustering and visualization for the term `Iran Election` (global concern). The properties of the users discussing this topic can be broken down demographically into four clusters.

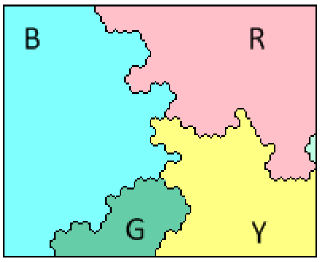


Figure 6.5: Final SOM-Ward clustering of metadata found in `Iran Election` tweets.

The blue area (detailed in Figure 6.6) represents users from various countries contributing to chatter about Iran's election aftermath. This user base is relatively new, predominantly Iranian web interface users participating on Twitter on the computer, with users only adopting Twitter for one month or less, and exhibits an emergent behavior of frequent reply-based messages [Cheong and Lee, 2010c,b].

The red area in Figure 6.6 is made up of almost mainly web users, from Iran and other countries; however this user base is more seasoned (or veteran) with accounts having been created at least three months prior. The contribution frequency is predominantly less than 10 messages per day, indicating sparing usage, but is contrasted by a high usage of other social media sites such as owning a blog or social network page[Cheong and Lee, 2010c,b].

The yellow area Figure 6.6 corresponds to adopters of social media who contribute to Twitter from Iran, the United States and the rest of the world. The inherent features of this cluster can be seen in the usage of mobile devices and social media applications, and the length of messages that hover among the 100-character range. I can deduce from this data that this segment of opinion holders is generating awareness of the Iranian situation via social media, possibly the younger generation [Cheong and Lee, 2010c,b].

The green cluster Figure 6.6 identifies users with little contribution rate (0-2 messages per day), but of varying Twitter account ages, and nationality. The properties emergent from this segment indicate a high posting of URL links in messages, and almost all of them using new categories of Twitter clients that are low in usage. The concentrated use of little-known Twitter clients suggest the depth of the Iranian elections topic in the sense that a very large spread of Twitter users participated in this topic of conversation [Cheong and Lee, 2010c,b].

Case 2: `iPhone`

The next case is to cluster the demographics for people using Twitter to express their thoughts about a consumer products (with a global scope), and where the area of discussion pertains to marketing and economics. My framework is tested on the trending topic keyword `iPhone` coinciding with the release of a new phone model by Apple Inc.

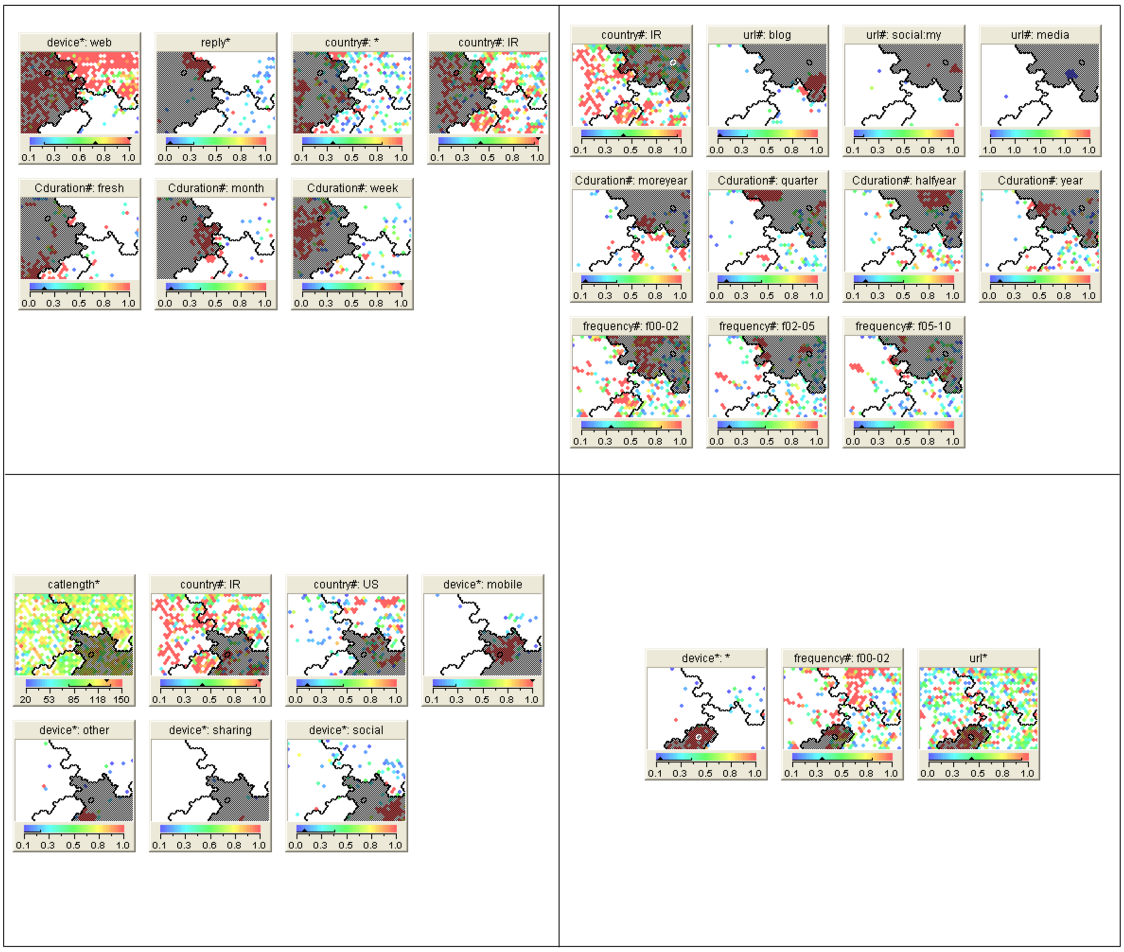


Figure 6.6: Emergent attributes for `Iran Election` tweets: (*top-left*) cluster 1/blue, (*top-right*) cluster 2/red, (*bottom-left*) cluster 3/yellow, (*bottom-right*) cluster 4/green.

Here, three distinct groups with several distinct emergent properties (Figure 6.7) were identified.

The blue cluster in Figure 6.8 identifies the source of the majority of chatter on Twitter regarding the iPhone. The demographics identified from this cluster are male, Twitter 'veterans' who have been adopting microblogging for at least a quarter of a year, but with an average daily contribution of less than five tweets. This user base comes from mainly Western countries where the *iPhone* has been marketed. This user base contributes to Twitter from a variety of devices (mobile and social network-enabled applications inclusive), and also have their own blog/website or social media site [Cheong and Lee, 2010c,b].

The second largest cluster in Figure 6.8 is colored red, whose emergent properties are accounts on Twitter that are significantly new (¡ 1 week), sourcing data from feeds such as RSS, have a significantly high ratio of followers to followees (higher in-degree than out-degree), and some contribute more than 50 posts per day on Twitter. Most of such messages have URLs in theme suggesting posting of links and content sharing by the users; and the majority of them have no country specified and no gender ascertained which suggest postings by news organizations or news aggregator sites. A small subset of this cluster which is of in-terest are messages belonging to Japanese (country code `.jp`),

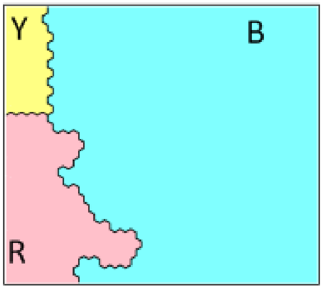


Figure 6.7: Final SOM-Ward clustering of metadata found in `iPhone` tweets.

which is notable as it reflects rather accurately the market sentiment of the *iPhone*'s new model launch in Japan [Martin, 2009; Cheong and Lee, 2010c,b].

The final cluster in Figure 6.8, which is the smallest, is colored yellow in the SOM above. The notable attributes and features for users of this cluster are that they are fresh accounts (one day old at the most), with unpopular social connections (*following* more people than they are friends with), with half of this cluster's Twitter accounts lacking in profile customization. They are posted predominantly from the web interface, frequently have more than 50 messages per day, and mention URLs in the links. As for demography, the gender and location information frequently could not be ascertained at all. I suspect that the Twitter chatter patterns for this group of users reflect those of opinion-spam and sockpuppetry (as hypothesized by Pang and Lee [2008]; Metaxas and Mustafaraj [2010]) who pollute the conversation stream with unnecessary noise: in other words, using Twitter for spamming and other disruptive purposes. A sample tweet in this cluster illustrates how such Twitter spam capitalizes on the popular nature of the *iPhone* topic [Cheong and Lee, 2010b]:

> "...[spam URL redacted] `Free iphone 's' I just got mine! where's yours!? Huh??`"

Case 3: `Obama`

The final case study conducted is a regional (mostly American) concern which delves into the realm of socio-politics, with a high-level SOM, visualized in Figure 6.9. The keyword `Obama` (for the US president) is tracked on Twitter to study the impact of his foreign policy statements (during time of writing of [Cheong and Lee, 2010c]) on Twitter user sentiment.

The biggest cluster of the user base detected in this study is the blue cluster in Figure 6.10, representing the demographics of Americans naturally concerned about the implications of Obama's foreign policy change. The user base contributes from a large array of devices (social-networking applications, mobile phones), but mainly via the web; and a substantial proportion of the users owning a website or blog. It is interesting to note that this cluster consists of users genuinely conversing about this topic, as their accounts are mainly more than 3 months old, their messages are almost always long, and their messaging style is focused to-wards replies, indicating conversation [Cheong and Lee, 2010c,b].

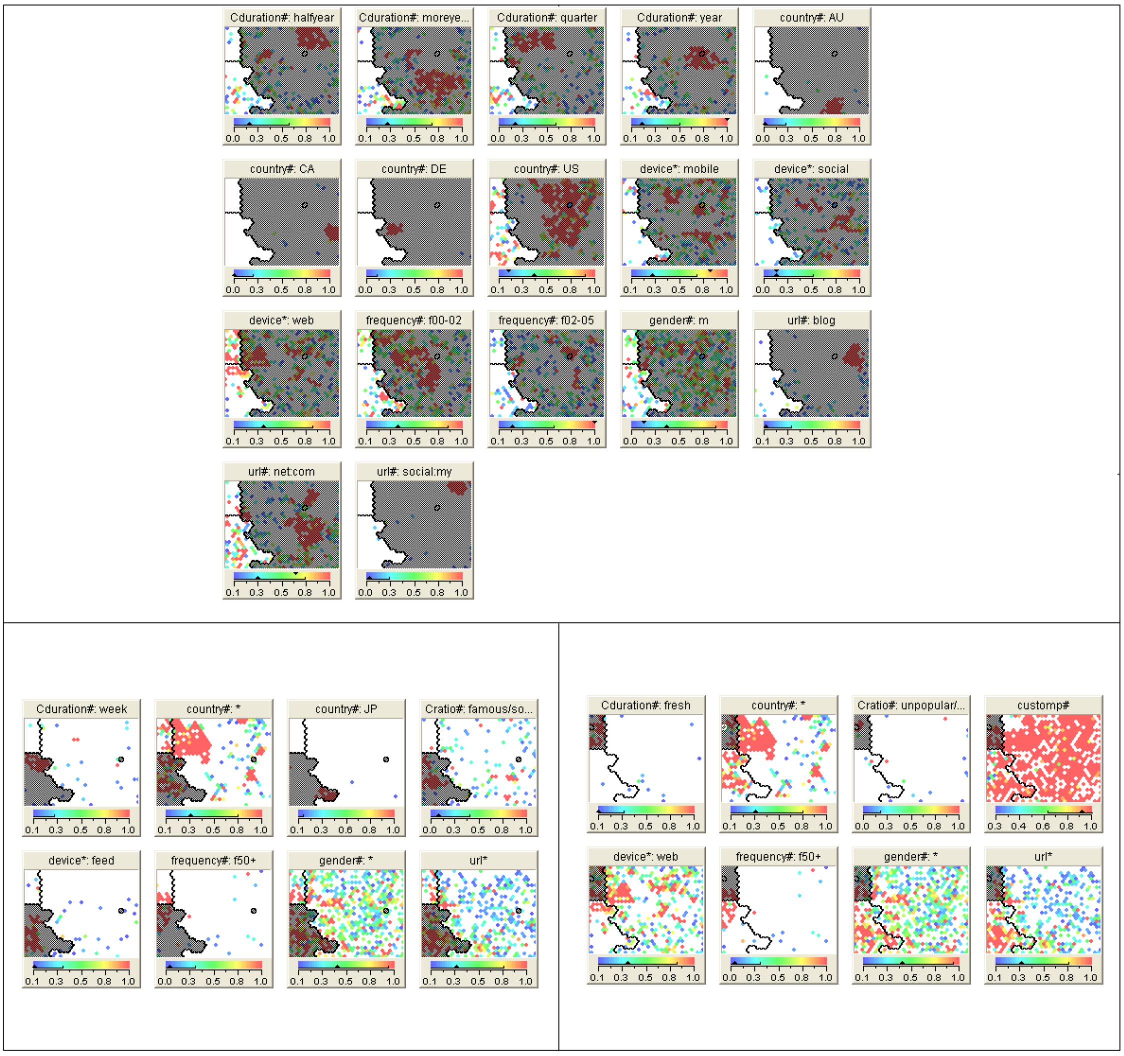


Figure 6.8: Emergent attributes for `iPhone` tweets: (*top*) cluster 1/blue, (*bottom-left*) cluster 2/red, (*bottom-right*) cluster 3/yellow.

The second largest cluster is colored red in Figure 6.10. I speculate that this cluster belongs to news sources and opinion leaders as the demographics reveal that users in this cluster have many followers (refuting the notion of opinion-spamming), predominantly US males, sourcing data from mostly data feeds such as RSS, and frequently publish URL links in their messages [Cheong and Lee, 2010c,b].

Finally, the yellow cluster in Figure 6.10 is composed of mainly new accounts, from indiscernible countries and genders, which mostly contribute postings from the web, leading me to suspect the presence of opinion-spam as discussed in the `iPhone` case study [Cheong and Lee, 2010b,c]. An example tweet in this cluster contains a number of hashtags (to gain visibility among users who are searching for particular tags) preceded by the text: "`the obamacare news center`." This is clearly an advertising tweet for a website that supports a particular view of Obama; which can be interpreted as a form of opinion-spam [Cheong and Lee, 2010b].

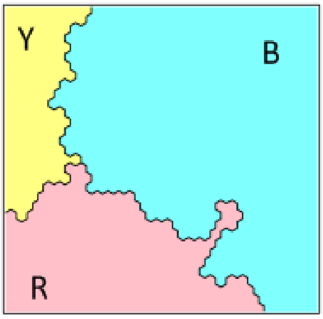


Figure 6.9: Final SOM-Ward clustering of metadata found in `Obama` tweets.

In conclusion, this case study has illustrated the richness in demographic data generated by automated inference algorithms. Clustering helps in pinpointing users from different socio-political topic areas in sentiment and opinion polling; and has potential in market segmentation and demographic segment targeting. The motivations and emergent properties of the users themselves is easily evident, as seen in Experiment 6.2, after the application of the SOM-Ward clustering algorithm.

Two minor follow-up studies have been conducted on the case study data set from above. First, a random survey on the spam clusters in the case studies above shows that the suspicious user profiles associated with the tweets have been removed by Twitter Inc. These results show promise in the efficacy of clustering in detecting spammers, complementing existing spam-detection approaches [Thomas et al., 2011; Metaxas and Mustafaraj, 2010; Lee et al., 2010]. Secondly, the visual representation of clustering can be used to visually inspect and validate the accuracy of clustering for the above case studies.

6.3.3 Study Conclusion

From Section 6.2 earlier, it was observed that SOM-Ward clustering worked well on features (such as gender and location) — deduced from user and message metadata — on a small-scale dataset (in the magnitude of ~100 records per topic).

The *Twitmographics* framework was able to automate the generation of inferences from raw data in the magnitude of thousands of records for a single topic. In the current section, through Experiment 6.2, I have determined that SOM-Ward was equally effective when it comes to clustering inferences that were obtained on a larger scale through the *Twitmographics* framework.

So far I have looked at how SOM-Ward clustering can be applied for pattern detection on the *Twitmographics Clustering Dataset* consisting of thousands of user and message records, as well as the smaller *SWSM2009 Clustering Dataset* which is a tenth of *Twitmographics Clustering Dataset*. In the following section, I will conduct a deeper investigation into the suitability of SOM-Ward clustering, as opposed to another popular, commonly-used, clustering algorithm — k-means — in relation to clustering Twitter inferences and metadata. A sample set of data will be extracted from the *10-Gigabyte Dataset* (Sections 4.5 and 5.3) and will be used as a testbed on which I can evaluate the results of SOM-Ward

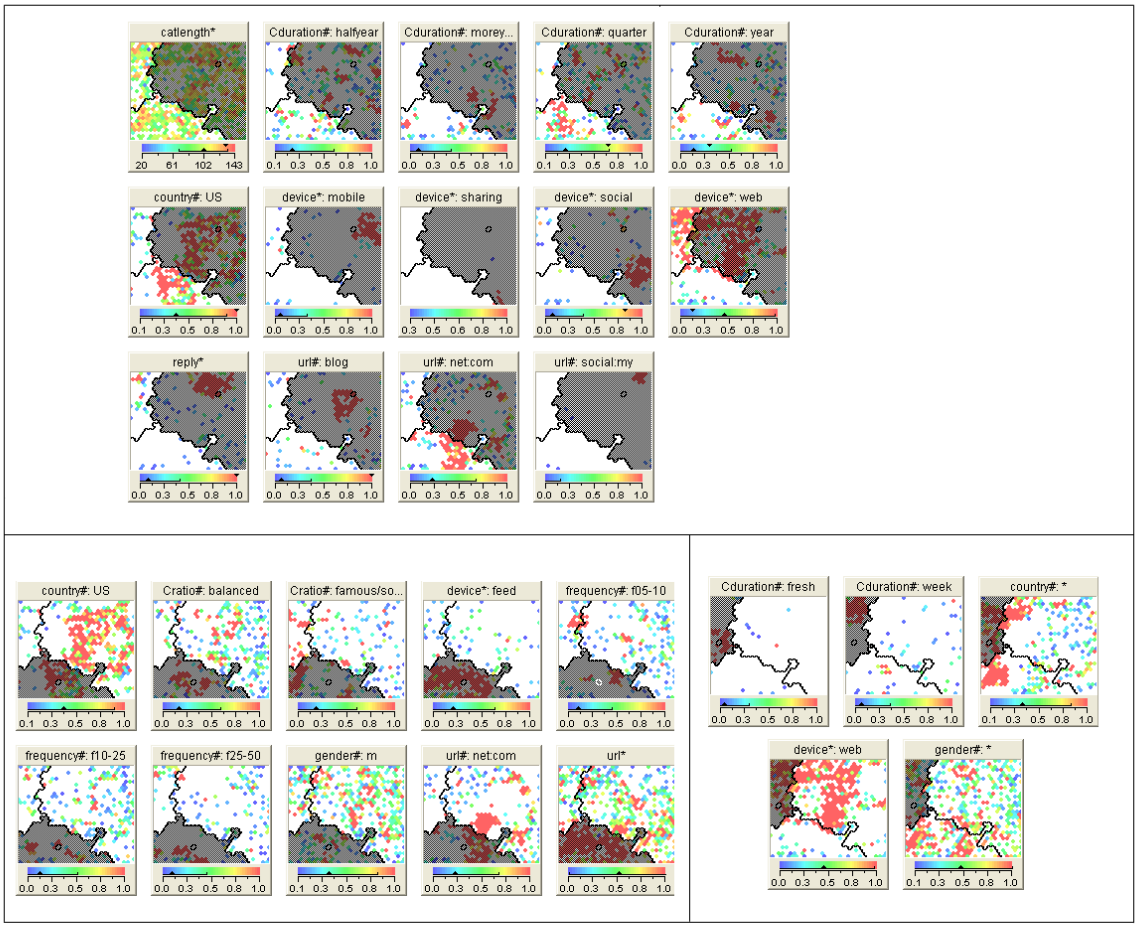


Figure 6.10: Emergent attributes for `Obama` tweets: (*top*) cluster 1/blue, (*bottom-left*) cluster 2/red, (*bottom-right*) cluster 3/yellow.

versus k-means clustering. A qualitative investigation into the features of the resultant clusters, as well as a quantitative evaluation in terms of internal validity of the clusters will also be examined.

6.4 SOM vs k-means Clustering on Twitter Metadata

In the previous sections, I have shown how SOM-based clustering has been shown in prior literature to be a suitable method [Couronné et al., 2009; Stoica et al., 2010; Torres, 2004; Merelo-Guervs et al., 2004; Yamazaki and Kumasaka, 2010; Claster et al., 2010] for clustering and visualizing resulting clusters of social network data.

The aim of this section is to perform a cursory comparison between clustering methods based on SOMs (such as SOM-Ward in Section 6.1.4) and the common k-means clustering method, introduced by MacQueen [1967].

I will provide an overview of the basic k-means algorithm; following that, I will touch on the theoretical differences between k-means and the SOM clustering method as highlighted in current literature. Lastly, I will evaluate the differences between SOM and k-means clustering on a real-world sample containing records with demographic properties and inferences obtained from Twitter messages, resulting from algorithms in Chapter 4.

6.4.1 An Overview of k-means

The k-means algorithm, also known as c-means or ISODATA, was first introduced by MacQueen [1967]. It is named as such based on the fact that a set of n data points are divided into k clusters. Each data point belongs to the cluster with a mean closest to itself [Mitra and Acharya, 2003; Jain and Dubes, 1988; Hartigan, 1975].

Advantages of k-means include the algorithm's simplicity, ease of implementation, ability to handle high-dimensional data, and its relative speed. However, one of its drawbacks is the fact that k-means is an NP-hard computational problem; this is remedied somewhat with modifications to k-means such as the incorporation of fuzziness and other heuristics.

Mathematically speaking, k-means partitions n data points into clusters
$S = \{S_1, S_2, ..., S_k\}$ such that the following condition is satisfied:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \tag{6.1}$$

given μ_i as the centroid (or mean of all the points) of a given cluster S_i.

6.4.2 Algorithm

A simple algorithm for k-means partitioning was proposed by Lloyd [1982]. Lloyd's Algorithm (the *de facto* k-means algorithm) iteratively performs the two main steps of — (a) recalculating the centroids or means for all k clusters, and (b) reassigning each data point to its nearest cluster. The algorithm converges when the members of each cluster no longer change [Jain and Dubes, 1988], as illustrated in Algorithm 6.10.

Algorithm 6.10 An algorithmic overview of Lloyd's Algorithm [Lloyd, 1982] for k-means clustering [MacQueen, 1967].

```
procedure KMeans
Require: set of $n$ data points to partition
    $k \leftarrow$ total number of clusters.
    for all clusters $S_i$, $i = \{1, ..., k\}$ do
        initialize centroid $\mu_i$ for each $S_i$
    end for
    repeat
        for all data points $x_j$, $j = \{1, ..., n\}$ do
            for all clusters $S_i$, $i = \{1, ..., k\}$ do
                calculate distance to centroid $d_{ij} = \|\mathbf{x}_j - \boldsymbol{\mu}_i\|$
            end for
            assign $x_j$ to cluster $S_i$, where $i, j$ satisfies $\min d_{ij} \in i = \{1, ..., k\}, j = \{1, ..., n\}$
        end for
        for all clusters $S_i$, $i = \{1, ..., k\}$ do
            recalculate centroid $\mu_i$ based on all members of $S_i$
        end for
    until cluster assignments cease to change
end procedure
```

6.4.3 Literature on k-means versus SOM-based clustering

In this chapter, I will report on existing findings in literature on the differences between SOM-based and k-means clustering techniques, in both theory and in practice on existing multidimensional data.

Theoretically, for the SOM algorithm, the SOM and k-means algorithms are “rigorously identical” [Bação et al., 2005] when the radius of the neighborhood function in the SOM equals zero [Bação et al., 2005; Chen et al., 2002]. In fact, a one-dimensional SOM — in which “...only the winner node’s weights are updated during [the training phase]” [Yao, 2006] — is equivalent to the online version of the k-means algorithm (online k-means is a version of k-means where the cluster centroids are updated as data points are assigned to clusters [Yao, 2006]).

Similar behavior is also expected if a SOM has a small number of nodes [Segev and Kantola, 2011].

Advantages of individual algorithms

In Abbas [2008], the author reports several general advantages of the k-means algorithm based on existing literature. The space complexity of k-means is light, at merely $O(k+n)$, excluding data matrix [Abbas, 2008]. Also, k-means is order-independent, as the same clustering of the data irrespective of the order in which the data is presented to the algorithm [Abbas, 2008].

Meanwhile, for SOM, the “combination of several map units allows [for] the construction of non-convex clusters”, which is impossible to perform in k-means clustering [Abbas, 2008]. Also, SOMs have been “successfully used for vector quantization and speech recognition” [Abbas, 2008].

Knowledge of initial cluster count

One of the issues faced with k-means clustering is that the algorithm (Algorithm 6.10) requires the value of k to be known in advance before the algorithm starts. The process of determining k is non-trivial [Chen et al., 2002; Abbas, 2008].

The same issue exists in determining the initial number of nodes in the lattice (or map size) of a SOM before running the algorithm [Chen et al., 2002; Abbas, 2008]. Again, the process of determining the optimal k is non-trivial. The examples of clustering illustrated in e.g. Sections 6.2 and 6.3 are based on an estimate of 10% of the total number of records to be clustered [Eudaptics Software GmbH, 2005]. As a general rule, however, the higher the value of k, the lower the performance of SOM-based clustering algorithms [Abbas, 2008].

Effects of sample size

For small datasets, SOM-based clustering algorithms show “good results” and are recommended for small datasets [Abbas, 2008]. On the other end of the spectrum, k-means is “very good” when using large datasets [Abbas, 2008].

Sensitivity to noise

With regards to noise in the input data — in the example of Twitter inferences, e.g. absent values for gender and location information — k-means and SOM-based clustering will exhibit differences in performance.

Abbas [2008] and Chen et al. [2002] documented that the k-means algorithm was "very sensitive" to noise in the dataset, which would "make it difficult... to cluster an object into its suitable cluster" [Abbas, 2008].

Structural quality and 'soft' clustering

Continuing from the topic of noise-sensitivity above, Chen et al. [2002] stated that the structural quality of SOMs is relatively low; i.e. the 'best fit' of the produced clusters w.r.t. the inherent partitioning of the data [Halkidi et al., 2001]. However, the neighborhood interaction was maintained, and the algorithm "gave relatively stable clusters" [Chen et al., 2002].

An advantage of SOMs, with respect to Twitter metadata, is that it rearranges data in a "fundamentally topological" order [Abbas, 2008; Chen et al., 2002; Segev and Kantola, 2011], "...correlated to the similarity of the clusters" [Chen et al., 2002].

In other words, it is easier "to observe relations between clusters" [Chen et al., 2002], valuable in achieving 'soft' clustering: i.e. "data [is] distributed diffusely and cannot be clearly segregated into isolated groups" [Chen et al., 2002]. For Twitter metadata and inferences generated from them, this property of a SOM is indeed beneficial as the features in Twitter datasets (see [Cheong and Lee, 2009, 2010c, 2011, 2010b]) oftentimes exhibit no clear intrinsic segmentation or partitioning.

Also, SOM is "less prone to local optima" [Bação et al., 2005] than k-means, as the search space is better explored by SOM [Bação et al., 2005]. k-means on the other hand "forces a premature convergence" [Bação et al., 2005]. Depending on how the clusters are initialized, k-means can often produce local optima can be achieved [Bação et al., 2005].

Note on clustering algorithms and implementations

In the comparative analyses of clustering algorithms by Abbas [2008], the author made a point about the virtually nil differences between different implementations (e.g. software/platform) of the same clustering algorithm. This is due to the fact that:

> "...the clustering algorithms using any software gives almost the same results even when changing any of the factors because most software use the same procedures and ideas in any algorithm implemented by them." [Abbas, 2008]

6.4.4 *Clustering Sample Dataset*: Twitter Data for Clustering Evaluation

The *10-Gigabyte Dataset* (discussed prior in Section 4.5 and 5.3) was used as the source of sample data. From the *10-Gigabyte Dataset*, I randomly sample 100 data records (containing both user and message metadata), and run my inference algorithms (Sections 4.6, 4.7, and 4.8) on these samples. For reference, the aforementioned dataset will be labeled as the *Clustering Sample Dataset*.

The types of tweets contained within the *Clustering Sample Dataset*, and by extension the properties of their authors, fall within five broad categories:

1. conversations and general chatter between users in English

2. conversations and general chatter between users in other languages

3. tweets in another character set (e.g. CJK characters: Chinese, Japanese, and Korean); such tweets contain a low message length (due to non-ASCII character santization), however they contain valid user properties and existence of message indicators such as hashtag and `@user` notations

4. chatter and retweets to (and of) celebrities, brands, and entertainment tweets

5. feed-generated or automated tweets (which includes spam)

Each of the data records contain a set of features, defined in Sections 4.6, 4.7, and 4.8, namely:

- **From the message domain**: presence of `@user` notation (nominal), presence of retweeting `RT` notation (nominal), presence of `#hashtags`, and device class (nominal).

- **From the user domain**: total post count (numeric), normalized activity frequency (numeric), follower/friend ratio (numeric), gender (nominal), country code (nominal), profile customization score (numeric), and profile website category (nominal).

The fact that my approach makes use of real-world everyday Twitter data — combining both nominal user metadata/inferences and statistical message properties — makes it difficult to create an intrinsic 'ground truth' of accurately-defined clusters in the *Clustering Sample Dataset*. It is important to note that the five categories above do not form a clear intrinsic segmentation or partitioning of the 100 data records, as my analyses involve the coupling of message properties together with user metadata and inferences. The best that can be done, in the case of the *Clustering Sample Dataset*, is the presence of "ideal types" [Milligan, 1996], which represents "an entity" (in the case of hundred samples: particular types of tweets and their authors) "...that will typify the characteristics of a cluster suspected to be present in the data" [Milligan, 1996].

Work has, however, indeed been performed on clustering similar tweets, e.g. [Horn, 2010; Ritter et al., 2010; Puniyani et al., 2010] (discussed in Sections 3.3.3 and 3.4.1). The difference is that such clustering experiments have a 'ground truth' of an intrinsic clustering of the data by means of human-assigned labels.

6.4.5 Proposed Validation of Clusters

There exist many techniques to analyze the validity of clusters in existing literature, divided into several broad categories:

- **Internal evaluation**: uses input data for clustering to assess quality of the clustering result, without depending on *external data* such as human-assigned class labels [Yao, 2006].
- **External evaluation**: comparing the clusters generated by a clustering algorithm "with an a priori partition of the data set" [Yao, 2006]. Simply put, the algorithm-generated clusters are compared to a ground truth of known cluster labels or external cluster assignments.
- **Relative validation**: a partition produced by a clustering algorithm is evaluated with respect to others "produced by the same algorithm, [initialized] with different parameters" [Yao, 2006].
- **Validation by visualization**: this is a simple approach where the clustering result is visualized and validated with a human experimenter [Yao, 2006]. This method does not work for large data sets or high dimensionality of features, and is rarely used for such reasons. However, this is still feasible by using SOM-based clustering methods, as SOMs project input from higher dimensions into maps of two-dimensions which can easily be visualized 6.1.2.

For this section, external evaluation is not usable, due to the absence of data labels and an intrinsic partitioning of the input data. Hence, for each of the clustering methods evaluated — k-means, and SOM-based Ward clustering (cf. Sections 6.2 and 6.3) — I perform a mixture of internal evaluation, relative validation, and visualization and interpretation of the results whenever possible.

A discussion on the various methods for internal evaluation of k-means clustering can be found in e.g. [Bezdek and Pal, 1998; Pal and Bezdek, 1995; Milligan, 1996]. For the purposes of internal evaluation in this section, I use the Davies-Bouldin index [Davies and Bouldin, 1979], which is a widely-used metric for internal evaluation of clustering results [Jain and Dubes, 1988].

Relative validation is used in evaluating k-means clustering, due to the fact that it is hard to determine an optimum value of k. Hence, I perform k-means clustering with differing values of k [Abbas, 2008] to determine the optimum number of clusters which yields the highest Davies-Bouldin index.

The Davies-Bouldin index is a measure which:

> "...indicates the similarity of clusters... can be used to infer the appropriateness of data partitions and can therefore be used to compare relative appropriateness of various divisions of the data. The measure does not depend on either the number of clusters analyzed nor the method of partitioning of the data" [Davies and Bouldin, 1979]

Given C_i as a cluster resulting from a clustering algorithm; X_j a data point (feature vector) in said cluster; and A_i the cluster centroid; S_i is obtained, which measures the scatter within the cluster:

$$S_i = \sqrt[q]{\frac{1}{T_i}\sum_{j=1}^{T_i} |X_j - A_i|^q} \tag{6.2}$$

given q as a constant (such that S_i is the "qth root of the qth moment of the points in cluster i about their mean" [Davies and Bouldin, 1979]). I now define $M_{i,j}$ as the measure of separation between C_i and C_j, also known as the Minkowski metric of centroids A_i and A_j [Davies and Bouldin, 1979]:

$$M_{i,j} = ||A_i - A_j||_p = \sqrt[p]{\sum_{k=1}^{n} |a_{k,i} - a_{k,j}|^p} \tag{6.3}$$

with a_{ki} being the kth component of the n-dimensional vector a_i (representing centroid i); and p a constant (a p value of 2 makes M_{ij} the Euclidean distance between A_i and A_j).

R_{ij} is, the ratio of S_i and M_{ij} that "reduces to certain familiar similarity measures for special choices of dispersion measures, distance measures, and characteristic vectors" [Davies and Bouldin, 1979].

$$R_{i,j} \equiv \frac{S_i + S_j}{M_{i,j}} \tag{6.4}$$

Finally, Davies-Bouldin index, $\bar{R}$, is obtained thusly:

$$\bar{R} \equiv \frac{1}{N}\sum_{i=1}^{N} R_i \tag{6.5}$$

given $R_i \equiv \max_{j:i\neq j} R_{i,j}$

Lastly, manual inspection and interpretation of the results of clustering are performed by using the generated cluster labels to annotate the original data points. Also, due to the inherent nature of SOM-Ward clustering, I am able to produce a graphical representation of the data points within SOM, which, to a certain extent, allows human interpretation of the results.

6.4.6 Evaluation of k-means Clustering on *Clustering Sample Dataset*

MATLAB (with the Statistics Toolbox) was used for k-means clustering. The SOM Toolbox [Alhoniemi et al., 2005], available from <`http://www.cis.hut.fi/somtoolbox/`>, is used for internal evaluation of the clustering results (via the `db_index()` function, an implementation of the Davies-Bouldin index) and also simple data normalization.

Before clustering can commence, I will need to determine the parameters for k-means clustering, especially the initial number of clusters, i.e. choice of k. Experiment 6.3 was designed to select the best value of k to be used for clustering the *Clustering Sample Dataset*.

Table 6.4: Parameters for k-means clustering in MATLAB's Statistics Toolbox used in evaluation.

Parameter	**Value**	**Description**
Distance metric	Squared Euclidean distance	Distance measure (between two data points) which is minimized during clustering. The squared Euclidean distance $d(x, y)$ is used; given two n-dimensional data points x and y, $d(x, y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2$.
Start algorithm	Cluster	Method used in obtaining the initial cluster centroids. The *cluster* option in MATLAB performs preliminary subsampling of 10% of the input data, and uses the obtained centroids to start clustering the entire data set.
Online phase	On	In addition to the 'batch update' phase which updates all cluster centroids in the naïve k-means algorithm, an online phase is also used where the centroids are updated 'online' as new data points are being examined.

Experiment 6.3. *Ascertaining the parameters for k-means clustering of the Clustering Sample Dataset by virtue of number of clusters and lowest Davies-Bouldin Index.*

METHOD: In clustering, the value of k needs to be known in advance; the process of determining k is non-trivial [Chen et al., 2002; Abbas, 2008] and is the subject of ongoing research. Therefore, in the vein of Abbas [2008], I performed clustering on a range of differing values for k, and observing the differences in the calculated Davies-Bouldin index.

The parameters for k-means clustering, implemented in MATLAB's Statistics Toolbox as the `kmeans()` function, are as per Table 6.4.

`kmeans()` was applied with the parameters above on the *Clustering Sample Dataset*, with values of $k = 2, 3, ..., 10$ (the maximum being 10% of the dataset sample size).

To account for the dominance of large values in features such as of number of posts, I use logarithmic normalization (also available as part of the SOM Toolbox in MATLAB) to preprocess the *Clustering Sample Dataset* before applying clustering. Simply put, this normalization method applies a log-transform to the data to reduce the influence of very large values in the dataset, which "is a good way to get more resolution to [the low end of a] vector component" [Alhoniemi et al., 2005]. As before, the resulting cluster assignments are re-evaluated using the Davies-Bouldin index.

RESULTS AND DISCUSSION: The k-means algorithm was first tested on the denormalized *Clustering Sample Dataset* (i.e. the data as-is) to observe its effects. On a denormalized sample, the clustering in general behaves well, in terms of the Davies-Bouldin Index, as shown in Table 6.4.

The value of $k = 2$ (italicized in Table 6.4) gives the lowest Davies-Bouldin Index for the denormalized dataset. This is due to the total number of posts by a given user — ranging from [20–7827] — dominating the clustering due to the absence of normalization.

Table 6.5: Evaluation of the Davies-Bouldin Index as a function of k, on the denormalized *Clustering Sample Dataset*. *Italics* indicate the k value responsible for the lowest index found.

k	**Davies-Bouldin Index**
2	*0.0699*
3	0.3971
4	0.4979
5	0.3847
6	0.4364
7	0.3993
8	0.3943
9	0.3107
10	0.4626

A cursory inspection of the clusters (where $k = 2$), with respect to the actual feature vectors of unnormalized *Clustering Sample Dataset*, reveals that all bar one of the sample data is clustered in one giant cluster. The second cluster contains only one member, which is an outlier in its own right. It turns out to be an automated Twitter account that publishes rental and house-sharing listings for Ontario, Canada. This is detected as an outlier due to the following:

- a very **high overall tweet count** (~187K);
- abnormally **high average daily activity** of 441.76 posts, which is very hard to achieve if the account is not automated (cf. Section 4.7.4);
- a **high FFR** of 527 usually only seen among famous people or information sources (cf. Section 4.7.3); and
- a **profile customization score of 2.0**, which is rare for automated/non-human Twitter accounts

From a statistical point of view, a clustering of $k = 2$ is optimal as it disregards the outlier, making this approach to clustering suitable in terms of outlier detection, as seen in e.g. Petrović [2006] in seeking anomalous patterns for intrusion detection. However, for the discovery of patterns such as those found in Cheong and Lee [2009, 2010b, 2011], such a clustering result does not suffice.

After normalization, i.e. log-transform on the data to reduce the influence of very large values in the dataset, the evaluation of cluster assignments using the Davies-Bouldin index are as per Table 6.6.

From Table 6.6, the lowest Davies-Bouldin Index results from $k = 10$ (0.8945). However, the high number of clusters will cause an interpretation of the cluster members to be highly specific (due to overfitting). Hence, I chose to analyze the partitions resulting from $k = 4$, which yielded the second-lowest Davies-Bouldin Index. Experiment 6.4 details the qualitative manual evaluation of cluster features.

Table 6.6: Evaluation of the Davies-Bouldin Index as a function of k, on the log-normalized *Clustering Sample Dataset*. *Italics* indicate the k value responsible for the two lowest indices found.

k	**Davies-Bouldin Index**
2	1.3973
3	1.0353
4	*0.9000*
5	1.1639
6	1.0214
7	0.9200
8	1.0146
9	1.0799
10	*0.8945*

Experiment 6.4. *Qualitative evaluation of the clustering of Clustering Sample Dataset, using the k-means algorithm ($k = 4$) in MATLAB's Statistics Toolbox.*

Method: From Experiment 6.3, I have ascertained that k-means clustering generates the best possible clustering on the log-normalized *Clustering Sample Dataset*. This is due to the low Davies-Bouldin Index obtained with a k value of 4, as per Table 6.6.

Using MATLAB's Statistics Toolbox, I used the findings from k-means clustering on the *Clustering Sample Dataset*, with $k = 4$, and all parameters fixed as per Table 6.4. I then manually inspected the features of the generated clusters, noting down any defining characteristics and peculiarities in the process.

Results and Discussion: Below are the defining characteristics of each of the four clusters as determined by the k-means algorithm, with a k of 4.

- **Cluster I**: This cluster contains the majority of the data points (67 out of 100 total records). The cluster contains a wide spread of message properties and user inferences. Almost 45% of the records originate from users who have detectable genders based on first names, which signifies that human users of Twitter constitute about half of this cluster (the rest consist of anonymous human users or automated Twitter accounts). About 15% of the records contain inferrable geographic information; while about 45% contain tweets which were composed on a mobile device. The distribution of usage statistics from user metadata, such as the FFR, total user post count, and messaging frequency (posts per day) are rather wide for this cluster.

- **Cluster II**: This cluster contains 3 data points, which are primarily characterized by a high user message count. Two of the tweets were composed by known automated Twitter programs (as seen in Section 4.6.3). All of the data points were composed by users with no readily-discernible demographic properties.

- **Cluster III**: The k-means algorithm successfully segmented out this cluster which contains 4 data points, all of which were composed by Twitter accounts with abnormally-high posts per day and FFR which are run by automated feed aggregators (software which collate RSS feeds from websites and reposts them as tweets).

- **Cluster IV**: The final cluster consists of 26 data points. From a human observer's perspective, the clustering algorithm successfully partitioned this cluster which primarily comprises of messages related to memes, particular celebrities, or pertaining to brands and marketing campaigns (see also Section 8.1 for a discussion on such trends). Empirically from the samples, I deduce that this cluster comprises mainly of human participants on Twitter taking part in trending behavior, due to the following indicators:
  - average FFR = ~2.41;
  - average user activity ratio = ~2.77 posts per day;
  - percentage of users with human names = ~61.54% (16 users), of which there are 10 females and 6 males; and
  - percentage of non-automated tweets = ~76.92% (20 tweets), including 12 mobile users of Twitter, 2 users of social media-integrated clients, and 6 Twitter website users.

6.4.7 Evaluation of SOM-Ward Clustering on *Clustering Sample Dataset*

As with evaluating k-means, MATLAB in conjunction with the SOM Toolbox [Alhoniemi et al., 2005] are used to evaluate the performance of SOM-Ward clustering. SOM-Ward clustering, as hitherto explored in Sections 6.2 and 6.3, is performed using the *Viscovery SOMine/Profiler* data-mining package. The clustering results are then ported to MATLAB to perform internal evaluation of the clustering results using the SOM Toolbox's inbuilt `db_index()` function. Experiment 6.5 documents the qualitative evaluation of SOM-Ward clustering on *Clustering Sample Dataset*, as well as a quantitative internal evaluation in terms of the Davies-Bouldin Index.

Experiment 6.5. *Qualitative evaluation of the clustering of Clustering Sample Dataset, using the SOM-Ward algorithm from Viscovery SOMine/Profiler, and internal evaluation of the clusters via the Davies-Bouldin Index.*

Method: The parameters in SOM construction and Ward clustering are chosen to resemble those used in prior studies in Sections 6.2 and 6.3, i.e. [Cheong and Lee, 2009, 2010b]. Compared to k-means, however, I do not have to initially determine the final number of clusters to be produced in partitioning. I will, however, need to determine the number of nodes in SOM construction: due to the small, easily-manageable size of the *Clustering Sample Dataset*, the number of nodes is set to the number of data points in the sample (100). Also, in SOM construction, I disable automatic correlation compensation, a feature in *Viscovery SOMine/Profiler* which is used to automatically reassign priorities to correlated attributes.

SOM-Ward clustering was performed on the *Clustering Sample Dataset*, where normalization is handled internally by the inbuilt algorithm found in *Viscovery SOMine/Profiler*. This resulted in the creation of 4 clusters, which is equal to the selection of k from the earlier evaluation of k-means clustering (Experiment 6.4).

Table 6.7: Parameters for SOM-Ward clustering of the *Clustering Sample Dataset* in *Viscovery SOMine*.

Parameter	**Value**	**Description**
Map Size	100	Number of nodes in the map, which is set to 100 units ≡ sample size.
Map Format	Automatic	Determines the aspect ratio of the resulting map; the *Automatic* option derives it from the "...ratio of the principal plane of the source data set." [Eudaptics Software GmbH, 2005]
Tension	0.5	Smaller tension values mean greater adaptation of the SOM to the data space ("the more the differences in the data are represented and the less the attribute values are averaged") [Eudaptics Software GmbH, 2005].
Training Schedule	Normal	Internally used by *Viscovery SOMine* to determine the speed, number of iterations and accuracy of the results; this ranges from *Fast* (least accurate, quickest), *Normal*, to *Accurate* (most accurate, slowest).

The cluster labels generated by *Viscovery SOMine/Profiler* were then imported into MATLAB. Using `db_index()`, the Davies-Bouldin Index of the resulting clusters will be calculated.

As per Experiment 6.4, I performed manual inspection of generated clusters and note any defining characteristics and peculiarities in the process. Compared to Experiment 6.4, in this experiment, the manual inspection was simplified by the use of visualization tools found in *Viscovery SOMine/Profiler* which allowed me to quickly examine features of interest from the generated maps.

Results and Discussion: Firstly, for internal evaluation, a Davies-Bouldin Index of 2.6414 was obtained from the clustering result. This relatively higher Davies-Bouldin Index compared to the one from k-means indicated that SOM-Ward clustering is less suitable than k-means, from a quantitative point of view, with respect to internal features of each cluster.

However, qualitatively, SOM-Ward clustering reveals a more intuitive clustering of the sample data; manual inspection of the data points in each cluster reveals the following distinguishing features between clusters:

- **Cluster I**: This cluster contains 42 out of 100 total records. As with Cluster I from k-means clustering, this cluster contains a wide spread of message properties and user inferences. Specifically for this SOM-Ward cluster, almost 74% of the records originate from users who have detectable genders based on first name; and 98% of tweets originate from a mobile device, web interface, or software client (as opposed to automated bots or feed harvesters). However, there is no readily discernible geographic information for users in this cluster. The distribution of usage statistics from user metadata, such as the FFR, total user post count, and messaging frequency (posts per day) are varied.

- **Cluster II**: This cluster contains 39 data points, almost all of which were composed by users with no readily-discernible demographic properties (gender nor location information). From manual inspection, less than half (46%) of this cluster comprise of foreign-language tweets, and almost all the tweets (save one) contain no `#hashtag` notation whatsoever.

- **Cluster III**: This cluster contains 10 data points, which are entirely comprised of automated (non-human) Twitter accounts with abnormally-high FFRs and total tweet count. All these accounts have `source` strings identifying the fact that they are run by automated feed aggregators.

- **Cluster IV**: The final cluster consists of 9 data points, comprising almost entirely by tweets authored by human users upon manual inspection. Demographic properties can be found in a majority of these records: 7 contain names that can be used to deduce gender, 8 records were authored using either a mobile device or the Twitter website, and all 9 records contain geographic information that was successfully geocoded. An interesting feature of this cluster is that 7 out of 9 records contain tweet content in a language other than English.

The features discussed with relation to the four clusters obtained with SOM-Ward can also be illustrated, due to the inherent nature of the SOM algorithm's 2-dimensional projection. Figure 6.11 highlights such features in terms of their individual feature maps.

In conclusion, based on experimental analyses on the *Clustering Sample Dataset*, k-means appears to have a better fit based on the Davies-Bouldin Index. However, qualitative empirical observation of the clusters indicate that k-means clustering poorly describes the features in Twitter inferences and metadata — ranging from large quantitative values such as FFR, to nominal values such as device classes — with no inherent structure.

Tou and González [1974] mentioned that clustering is "very much an experiment-oriented 'art'...", due to the fact that it depends on "the type of data being analyzed" and is heavily influenced by "the chosed measure of pattern similarity and the method used for identifying clusters" [Tou and González, 1974]. With this in mind, I have found that SOM-Ward, can indeed be used effectively to discover interesting topological patterns with respect to Twitter inferences. This was despite SOM-Ward's higher Davies-Bouldin Index as tested on *Clustering Sample Dataset*, and clusters not as tight as the ones generated with k-means [Chen et al., 2002].

6.4.8 Concluding Notes

This chapter has covered four key areas in terms of clustering Twitter metadata and metadata-based inferences. In Section 6.1, I have discussed how SOMs have been useful for the clustering on social network data, which Twitter is a form of. The SOM-Ward algorithm, which combines the map generation by Kohonen [1988] and clustering by Ward [1963], was elaborated upon, in terms of the key concepts involved, and also its advantages in clustering social network data.

Section 6.2 details my first foray into the application of clustering for the detection of latent features within a Twitter dataset, viz. *SWSM2009 Clustering Dataset*. Qualitative evaluations of the results has shown the effectiveness of SOM-Ward in producing meaningful segmentations of the users (and their tweets). Section 6.3 builds from the prior section as it deals with a significantly larger dataset, *Twitmographics Clustering Dataset*, generated using automated approaches (Sections 4.6, 4.7, and 4.8). Again, SOM-Ward clustering has proven useful in this endeavor, despite the larger feature space and sample size compared to manual inspection of tweets.

Finally, Section 6.4 compared SOM-Ward clustering with k-means clustering, which is another popular approach used in pattern recognition. I evaluated the results of both clustering methods on the *Clustering Sample Dataset* dataset, which was built for the sole purpose of comparing clustering methods. The resultant clusters from both SOM-Ward and k-means were evaluated quantitatively using the Davies-Bouldin Index, an internal evaluation metric to quantitatively assess the quality of the clustering result, and also qualitatively via the inspection of features and data points in each cluster.

Chapter 7, which follows, will consist of the applications of the approaches seen in Chapters 4, 5, and 6 on Twitter data pertaining to real-world phenomena. The three case studies in Chapter 7 will serve to illustrate how the techniques, algorithms, and findings from the aforementioned chapters fit together in the analysis of Twitter data from real users, resulting from said real-world events.

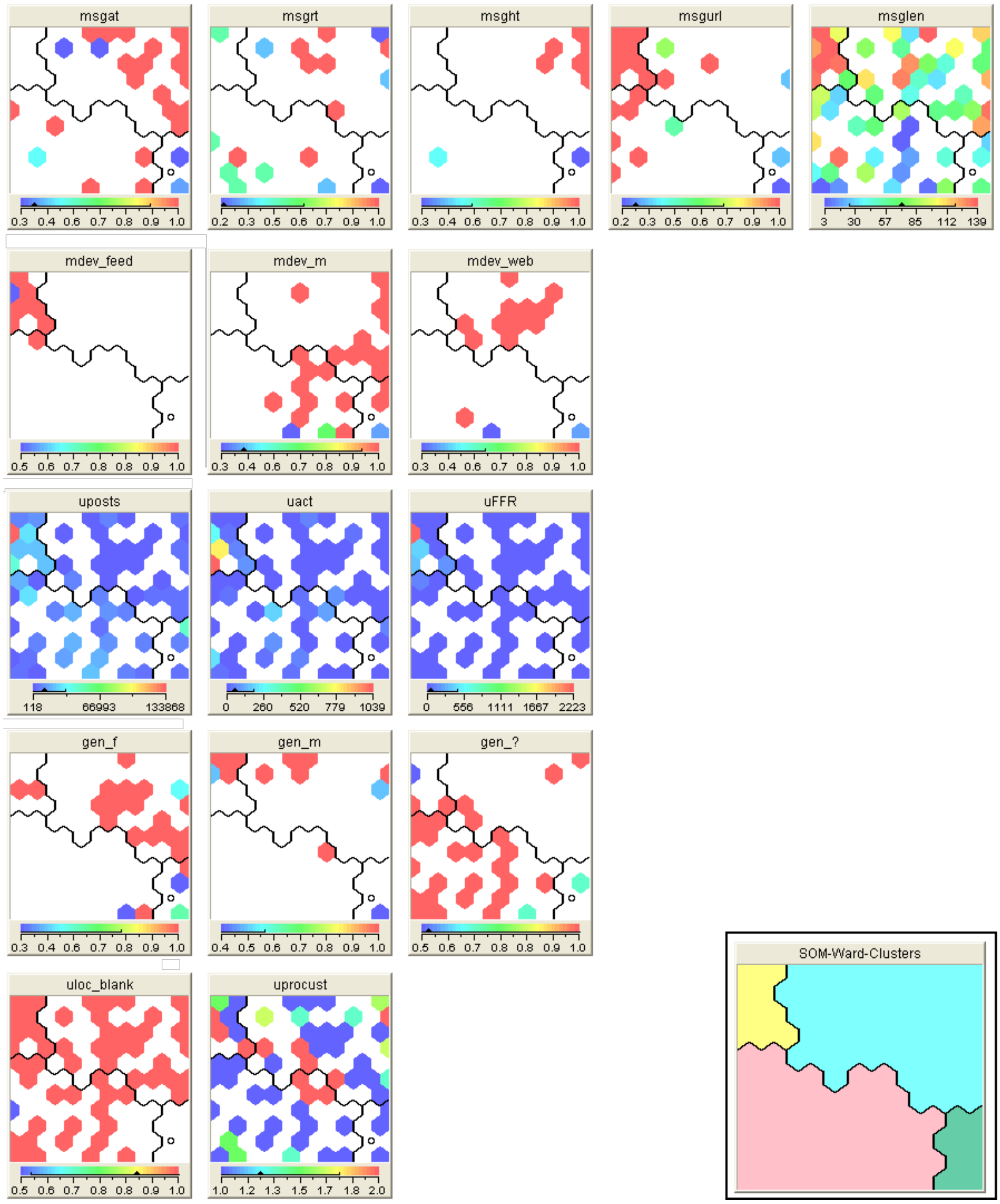


Figure 6.11: Result of SOM-Ward clustering of metadata for the *Clustering Sample Dataset*, with the overall map (inset, bottom-right), and maps of individual features that are of interest.
Row 1 (left-to-right): `@user` notation; retweet presence; `#hashtag` presence; URL presence; and message length.
Row 2 (left-to-right): (binarized nominal attributes for device classes) feed aggregator; mobile client; and Twitter web interface.
Row 3 (left-to-right): total user tweet count; normalized user daily tweet frequency; user FFR.
Row 4 (left-to-right): (binarized nominal attributes for user gender) male; female; indeterminate.
Row 5 (left-to-right): map showing absence of geocodable location information; profile customization score.

Chapter 7

Case Studies: Real-World Events Seen via Twitter

> "*Is this the real life? Is this just fantasy?*
> *Caught in a landslide; no escape from reality.*
> *Open your eyes; look up to the skies and see...*"
>
> — Queen
> *The Bohemian Rhapsody* (1975)

Parts of this chapter have been published as:

Cheong, M. and Lee, V. [2010d]. Twittering for Earth: A Study on the Impact of Microblogging Activism on Earth Hour 2009 in Australia, *Proc. ACIIDS 2010.*

Cheong, M. and Lee, V. [2011]. A Microblogging-based Approach to Terrorism Informatics: Exploration and Chronicling Civilian Sentiment and Response to Terrorism Events via Twitter, Information Systems Frontiers **13**(1): 45–59.

Cheong, M., Ray, S. and Green, D. [2012a]. Interpreting the 2011 London Riots from Twitter Metadata, *Proc. SoCPAR 2012.*

In Chapter 4, I have detailed the kinds of user and message metadata available from Twitter APIs, and how the said metadata on Twitter can be converted into real-world inferences using algorithms developed as part of my research. Consequently, in Chapter 5, I have detailed two frameworks for automated collection of metadata from Twitter via its APIs, both on-demand and streaming, and applied the algorithms (introduced in Chapter 4) on the large *10-Gigabyte Dataset* of real life tweets and users. This segues into Chapter 6 which examines how meaningful patterns can be obtained from metadata-based inferences when pattern-recognition and clustering techniques (such as SOM and k-means) are applied.

Having researched on such new methods and applications of pattern recognition on Twitter inferences, in both the user and message domains, I will now demonstrate their efficacy to deal with real cases on Twitter. I investigated three cases, designed to illustrate a range of issues concerning evaluation of real-world tweets (and the users) behind such phenomena.

I will firstly elaborate on a four-year longitudinal study on Twitter activism. This study monitored Twitter for Australian user participation in the Earth Hour campaign through four years: 2009–2012. I have captured tweets by Australian Twitter users during Earth Hour on a state-by-state basis. Using per-message entity analysis (Section 4.8.2), I measured the effectiveness of the Earth Hour campaign, by virtue of yearly energy savings recorded during Earth Hour versus yearly Earth Hour Twitter activity. As an aside, I also attempted to determine a link between Twitter participation as a function of state-by-state population. This study fits in with the theme of this chapter by demonstrating how Twitter metadata inferences can be used to fathom real-world activism campaigns.

The second part of this chapter focuses on how Twitter can potentially be used to chronicle useful information during terrorism events. This part ties in with the overall aim of this chapter as such events are real-world events that have a significant presence on Twitter, as illustrated by e.g. the Mumbai [Beaumont, 2008] and Jakarta attacks [Saputra and Leitsinger, 2009]. By adapting my Twitter data-harvesting framework (Section 5.1), I was able to craft a prototype framework for Twitter-based terrorism informatics that could be of use to law enforcement and academia. This leverages on the basic idea of deriving important user and message inferences (Chapter 4), augmented with useful observations from extant literature on civilian behavior during terror events. I will also discuss potential methods of visualizing terrorism events, by applying 2D visualization and visual clustering algorithms (Chapter 6) on the user and message inferences.

The final part of this chapter is where I document a case study of how user and message metadata can be combined to give a birds-eye view of the 2012 London Riots, as documented via Twitter. The demographics and communication properties of Twitter users involved in the riot will be exposed using inference algorithms (Chapter 4). The inferences generated are also used in combination with pattern recognition techniques (Chapter 6) to reveal latent patterns of rioters and other people discussing the riots in the Twitterverse.

7.1 ‘Twittering For Earth’: Australian Earth Hour Activism

This case study, first published as [Cheong and Lee, 2010d], illustrates how online activity on Twitter which is part of the “social web” [Cheong and Lee, 2009] can be reflected in a real-world social system. Such studies provide insight as to how social media in a virtual online setting can be linked to real-world human behavior, which supplements existing studies on online memetics [Arbesman, 2004; Wasik, 2009], information-sharing behavior [Cheong, 2009], and social participation and dynamics.

In the original published study [Cheong and Lee, 2010d], I investigate how Twitter activity in a collective action campaign can be a reflective indicator of real-world sentiment on real-world events — in this case, the microblogging pattern of Australian Earth Hour 2009 participants on Twitter. This case study, which has since been expanded to include observations from the 2010, 2011, and 2012 editions of Earth Hour, has four goals:

- **Study Goal 1**: Measuring the difference in energy consumption during Earth Hour with respect to average power consumption levels for each of the 2009–2012 Earth Hours.

- **Study Goal 2**: Measuring the state-by-state Twitter participation of the 2009–2012 Australian Earth Hour campaigns, studying the trends across the four years, and the potential causes behind such changes.

- **Study Goal 3**: Discovering any correlation between per-state real-world energy savings with respect to online Twitter activity (which shows how Twitter can be used as a 'mirror' for real world events.

- **Study Goal 4**: Finally, as a side investigation, discovering any correlation between per-state Twitter activity with state population (to see if adoption rates of online microblogging and social networking technologies can be linked to the size of the population, in the Earth Hour context).

As far as I know, no prior work has been done with respect to this topic. The potential results obtained from such a study could illustrate how an online microblogging platform — with elements of social networking and communication [Cheong and Lee, 2009; Honeycutt and Herring, 2009], collaborative applications [Honeycutt and Herring, 2009], and information dissemination [Cheong, 2009; Java et al., 2009; Mischaud, 2007]) — could 'mirror' a real-world social system effectively.

7.1.1 Background

Studies in activism among users of *de facto* online social networks (OSNs) are common in research, particularly in the domain of social sciences and the humanities. A study by Song [2008] in feminist cyber-activism showed how Facebook — an OSN — can be leveraged to enhance activism. Mankoff et al. [2007] have also conducted a study of how such OSNs can be used in eco-activism, by encouraging members on OSNs to reduce their ecological footprint in real life.

Relating to this case study, there was prior work done by Solomon [2008] on analyzing the energy drop recorded during the 2007 Earth Hour conducted Sydney-wide from an economic perspective, which revealed findings that users "overstate their participation in the Earth Hour project", as observed from the total energy drop registered during the 2007 Earth Hour. This study also investigates whether the energy savings have significantly changed from year-to-year (cf. *Study Goal 1* of this section), albeit involving data which is rather dated.

7.1.2 Isolating `#earthhour` Twitter Chatter

For the Earth Hour campaigns, users on Twitter are encouraged by the organizers to publish Twitter messages in such a format to express their support for Earth Hour:

> *. . . use the hashtags* `#earthhour` *or* `#voteearth` *along with your* `#location` *to get the word out.* [William-Ross, 2009]

Hashtags illustrate the "use of social tagging to categorize posts" [Cheong, 2009] allowing for organization and simplify searching of related posts [Huang et al., 2010; Makice, 2009a].

Using the specific hashtag notation as per William-Ross [2009], I was able to craft a search query to seek out Twitter messages from Australia in support of Earth Hour. Generally, my data collection methodology remains the same throughout the four years of observation. However, several technical changes have taken place due to the evolution of the Twitter Streaming API (Section 4.2) and the deprecation of the previously-used on-demand APIs (Section 4.1.3).

7.1.3 Tweet Collection

2009 Data Collection

Initially, in 2009, I planned to use the Twitter on-demand APIs — `search` and REST-`user` (Section 4.1) — to collect Earth Hour tweets. However, due to timing issues, I could not use the on-demand APIs for data collection, as it imposed a limit on the number of backdated results. The back-date period was approximately one month [Cheong and Lee, 2009], which is insufficient as my experiment was designed a few months after the Earth Hour event. Therefore for the purposes of obtaining Earth Hour 2009 data, I use the *Hashtags.org* website [Bailey et al., 2009] as the data source. *Hashtags.org* is a Twitter REST API-based website which automatically tracks users with hashtags and has a backdated, browsable archive for the hashtags '`#earthhour`' or '`#voteearth`'.

2010 Data Collection

For the 2010 edition, to correct for the constraint in obtaining historical search results, I have improvised the use of a simple Perl script that polls the Twitter `search` API (as a on-demand/REST service) at set intervals, and timed to run during the Earth Hour event. This generates a near-continuous 'live' stream of messages as Earth Hour was taking place. A continuous stream of the messages on Twitter via the (then-fledgling) Streaming API would indeed be the optimal solution. However, as the Streaming API was still undergoing development in 2010, and the low volume of messages available to researchers during the time period, it was not feasible for data collection as of the 2010 Earth Hour event.

2011-2012 Data Collection

As development has matured, the Streaming API has become more suitable for large-scale data collection circa 2011 (Section 4.2.3). Therefore, Twitter data for 2011 and 2012 Earth

Hours were collected by listening to the Streaming API using a *TweetHarvester* prototype (Section 5.2).

7.1.4 Location Filtering via Hashtags

To obtain the location wherein an Earth Hour tweet is authored, the list of Earth Hour messages obtained are simply iterated to identify entities (Section 4.8.2) containing Australian state capitals, major cities, and their abbreviations. The messages are then collated according to state (with case-insensitive matching), as per the hashtag keywords in Table 7.1. This was in line with the original location-tagging concept espoused by the Earth Hour organizers [William-Ross, 2009]; additional metadata inference techniques (e.g. Section 4.6.2) were not applied.

Table 7.1: Hashtag keywords found in Earth Hour Twitter messages Australia-wide, grouped on a state-by-state basis.

State	**String describing state name, abbreviation, state capital, or major cities**
NSW	`Nsw, NewSouthWales, Sydney`
QLD	`Qld, Queensland, Brisbane`
SA	`SouthAustralia, Adelaide`
TAS	`Tasmania, Hobart`
VIC	`Victoria, Melbourne, Bendigo`

From the corpus of the filtered Twitter messages, a measure of *Twitter activity* is obtained: i.e. total Twitter messages as per each of the aforementioned states in Table 7.1.

7.1.5 Power Consumption: Acquisition and Analysis

The Australian National Electricity Market Management Company Limited (NEMMCO) [National Electricity Market Management Company Limited, 2009][1] publishes electricity market (supply and demand) data, updated on a half-hourly basis, for five Australian states: New South Wales, Queensland, South Australia, Tasmania, and Victoria. Energy market data comes from the website of the National Electricity Market Management Company Limited [2009], which publishes electricity market supply-and-demand data for the five Australian states on a half-hourly basis. With this data, I was able to come up with an authoritative measure of how much energy is used; from which I can estimate the energy saved during the 2009 observance of Earth Hour (8.30pm to 9.30pm local time for each state).

As of July 2009, NEMMCO's operations have been taken over by the Australian Energy Market Operator (AEMO), which still provides the same data for the analysis of the 2010–2012 Earth Hour campaign's power consumption on their website [Australian Energy Market Operator, 2012].

[1] As of July 2009, NEMMCO's operations have been taken over by the Australian Energy Market Operator (AEMO).

Power consumption data in megawatts (MW) are obtained for the Earth Hour time period (the two half-hour periods of 8.30pm–9.00pm and 9.00pm–9.30pm) and the corresponding power consumption data for three days before surrounding Earth Hour (as the baseline). The dates included in this observation period are detailed in Table 7.2.

Year	**Earth Hour day**	**Monitoring period**
2009	Saturday 28 March	25 March–31 March inclusive
2010	Saturday 27 March	24 March–30 March inclusive
2011	Saturday 26 March	23 March–29 March inclusive
2012	Saturday 31 March	28 March–3 April inclusive

Table 7.2: The day on which Earth Hour falls on, for the years 2009–2012, and the seven-day period where average wattage is monitored.

The average non-Earth Hour power consumption (on a state-by-state basis) is calculated by averaging the wattage for the three days before and three days after the Earth Hour event. Hence, the energy reduction during Earth Hour could be expressed as a percentage of the average consumption.

To test the significance of the energy drop, a paired Student's t-Test for statistical significance is performed (with significance level α=0.05) using a spreadsheet package on the entire set of wattage data, to see if the savings in power is significant enough statistically.

Student's paired t-Test [Student, 1908] is used on two dependent or 'matched' samples, in order to test the null hypothesis of both samples having the same mean (i.e. the difference between the sample means is zero). It makes use of the t-value, which is defined mathematically as:

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}} \tag{7.1}$$

where $\overline{X}_D$ is the mean of the differences between the two samples, and s_D is the standard deviation of the differences between the two samples. The null hypothesis is $\mu_0 = 0$; i.e. zero difference between the sample means. n is the number of values in each sample.

7.1.6 Experimental Analyses

This section details the experiments involved in achieving *Study Goals 1–5*, and discusses their results.

Experiment 7.1. *To measure difference in energy consumption during Earth Hour (Study Goal 1).*

METHOD: For each of the surveyed Earth Hours, I calculate the average state-by-state consumption of power in three days before and after Earth Hour, for two half-hourly periods in each state's respective time zone (as previously mentioned in Table 7.2).

The periods are defined as:

- **Period 1**: 8.30pm — 9.00pm

- **Period 2**: 9.00pm — 9.30pm

Table 7.4 lists the per-state average energy consumption (three days before and after the Earth Hour for a given year), and Earth Hour energy consumption, for the period 2009-2012.

Using a spreadsheet, the t-probability using Student’s two-tailed, paired t-Test for the two samples (the average consumption data and corresponding Earth Hour consumption data for a given year) is calculated. In this experiment, for each year, the null hypothesis is *a zero difference between the means of both samples.*

For each year’s results, the t-probability was obtained using a two-tailed, paired t-test. Given the t-probability, I obtain the inverse of Student’s t cumulative distribution function (c.d.f.) with $\alpha = 0.05$, and degrees of freedom (d.o.f.) $= 18$ (Equation 7.1).

The obtained inverse-t values for the yearly data as well as its corresponding t-probabilities are listed in Table 7.3. From the given parameters, an inverse-t value of 2.10 or more is required to indicate a statistically significant drop of energy consumption during Earth Hour based on Student’s t-Test (and hence refute the null hypothesis).

Results and Discussion: Each of the yearly inverse-t values obtained in Table 7.3 exceed 2.10, indicating that the energy drops for the four years are statistically significant. The null hypothesis $\mu_0 = 0$ — i.e. no difference between the means of the two samples — is thus refuted for each year.

Table 7.3: t-values obtained using Student’s two-tailed, paired t-Test on energy reduction levels for Earth Hour; with $\alpha = 0.05$, degrees of freedom = 18.

Year	**t-probability from a paired, two-tailed t-test**	**inverse-t c.d.f. with $\alpha = 0.05$, d.o.f. $= 18$**
2009	0.0046	3.24
2010	0.0028	3.46
2011	0.0007	4.09
2012	0.0008	4.02

Table 7.4: State-by-state energy consumption for two 30-minute periods during Earth Hour, and the average consumption for the same periods on non-Earth Hour days (in Megawatts), recorded for the Earth Hour events of 2009–2012 [National Electricity Market Management Company Limited, 2009; Australian Energy Market Operator, 2012].

2009										
State & period	**NSW:1**	**NSW:2**	**QLD:1**	**QLD:2**	**SA:1**	**SA:2**	**TAS:1**	**TAS:2**	**VIC:1**	**VIC:2**
Average wattage	8688.59	8373.54	6454.31	6330.06	1496.56	1450.41	1071.12	1030.07	5806.23	5659.26
Earth Hour wattage	7729.56	7607.68	6001.81	5789.71	1370.52	1372.22	1031.49	1008.6	5317.49	5323.03
Difference	959.03	765.86	452.5	540.35	126.04	78.19	39.63	21.47	488.74	336.23

2010										
State & period	**NSW:1**	**NSW:2**	**QLD:1**	**QLD:2**	**SA:1**	**SA:2**	**TAS:1**	**TAS:2**	**VIC:1**	**VIC:2**
Average wattage	9103.65	8786.86	6498.03	6374.98	1552.16	1513.06	1160.62	1132.48	5846.72	5676.09
Earth Hour wattage	8576.47	8463.09	6087.91	5915.33	1444.95	1412.96	1182.02	1158.43	5390.64	5307.20
Difference	527.18	323.77	410.12	459.65	107.21	100.10	-21.40	-25.95	456.08	368.89

2011										
State & period	**NSW:1**	**NSW:2**	**QLD:1**	**QLD:2**	**SA:1**	**SA:2**	**TAS:1**	**TAS:2**	**VIC:1**	**VIC:2**
Average wattage	8635.24	8378.34	6452.91	6293.81	1526.01	1498.62	1072.94	1045.32	5690.61	5518.61
Earth Hour wattage	7805.38	7725.25	5956.14	5836.63	1296.95	1271.88	978.36	970.92	5265.69	5154.19
Difference	829.86	653.09	496.77	457.18	229.06	226.74	94.58	74.40	424.92	364.42

2012										
State & period	**NSW:1**	**NSW:2**	**QLD:1**	**QLD:2**	**SA:1**	**SA:2**	**TAS:1**	**TAS:2**	**VIC:1**	**VIC:2**
Average wattage	8373.22	8144.72	6115.41	6026.48	1630.37	1596.67	1090.50	1064.09	5740.73	5569.27
Earth Hour wattage	7602.82	7465.35	5673.38	5598.32	1483.45	1447.82	982.36	960.81	5123.53	5008.32
Difference	770.40	679.37	442.03	428.16	146.92	148.85	108.14	103.28	617.20	560.95

Figure 7.1 visualizes the yearly energy savings for each of the states studied in Experiment 7.1. Each state’s annual energy savings during Earth Hour is expressed as a percentage of the average consumption for that given state.

For this experiment, every state recorded energy savings for Earth Hour every year, except for Tasmania which recorded negative energy savings (excessive consumption) in the 2010 Earth Hour.

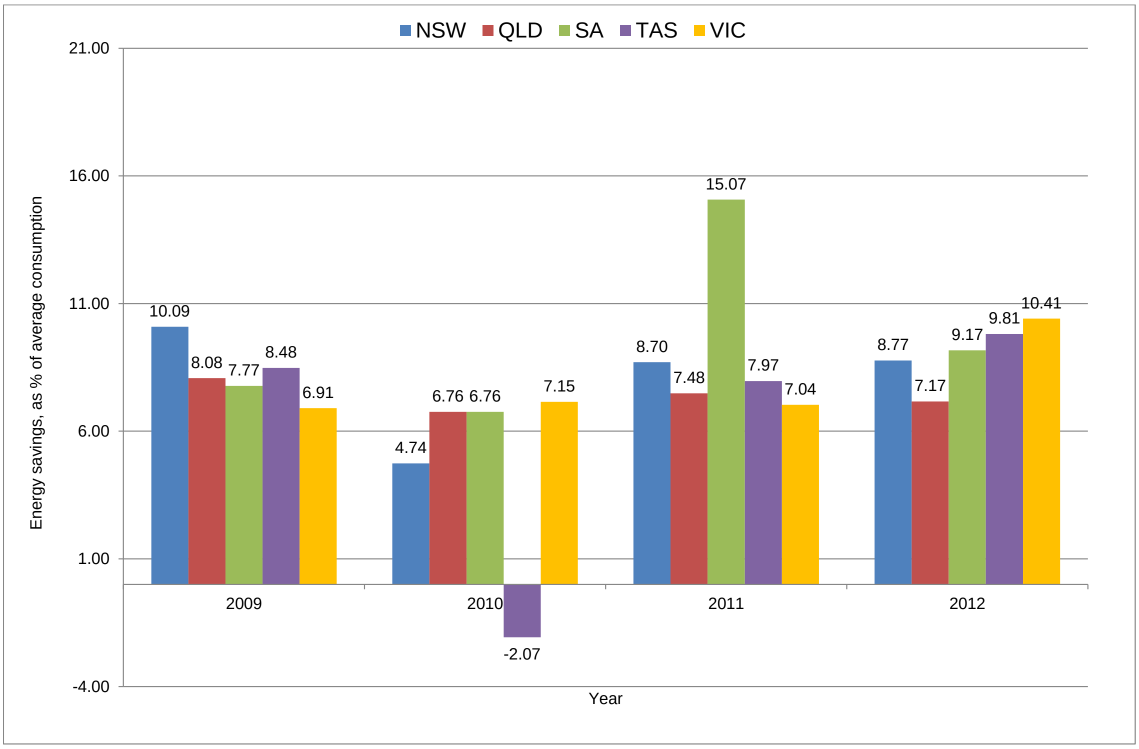


Figure 7.1: Comparison of energy saved (as a percentage of average consumption in megawatts), across the Earth Hours 2009–2012 [National Electricity Market Management Company Limited, 2009; Australian Energy Market Operator, 2012].

Based on the total energy savings recorded in this experiment, the drop of energy consumption suggests statistically significant results in promoting energy conservation across the 2009-2012 Earth Hours. Although this study is confined to five major Australian states — NSW, QLD, SA, TAS and VIC — it does however suggest an efficacy on the part of the Earth Hour organizers in promoting awareness of energy conservation.

Among the chief reasons for this is the ‘buzz’ generated among the populace by a successful marketing campaign involving commitments from local governments worldwide, and the engagement of social media (blogs, social networks, and microblogging sites). These corroborate research illustrating the efficacy of social media in activism and also the viral spread of information online [Arbesman, 2004; Wasik, 2009; van Liere, 2010].

In 2010, however, despite having statistically significant energy savings, the average savings percentage has decreased since 2009. I observed a lack of participation for the 2010 Earth Hour, as indicative of tweets in the 2010 Earth Hour such as:

- `Some huge building in the city isn’t switching off any of their lights :(`

- `it strikes me that #earthhour means that the national grid has to waste a whole load of energy to the atmosphere. #fail`

In 2011 and 2012, the energy savings are still statistically significant as per the *t*-Test parameters (Table 7.3). The average percentage of energy savings, as a whole, is slightly better compared to the results in 2009–2010.

Experiment 7.2. *Measuring state-by-state Twitter participation of the 2009–2012 Australian Earth Hour campaigns (Study Goal 2).*

METHOD: The *Twitter participation of users* in each state is simply the number of Earth Hour Twitter messages referring to a particular state. From the set of filtered Twitter messages (as elaborated in Section 7.1.4), I determine the number of tweets linked to individual states via hashtags in Table 7.1..

RESULTS AND DISCUSSION: Table 7.5 lists the tally of messages from all five states, across four years[2].

Table 7.5: Total tally of Earth Hour Twitter messages across five years, from each state.

State	2009	2010	2011	2012
NSW	48	16	52	27
QLD	21	4	22	5
SA	6	8	4	8
TAS	3	0	17	0
VIC	24	9	34	9
Yearly total	**102**	**37**	**129**	**48**

This figure varies from year-to-year due to several reasons, three of which I have identified as follows:

1. **API differences:** Firstly, the different APIs used, and their associated limits on message quota, affect the size of the overall sample space from which the messages are extracted. The `search` API used in the 2010 experiment was able to provide results from the entire collection of Twitter messages, albeit having severe limits on retrieval. On the other hand, the Streaming API used in the 2011 and 2012 experiments provides a large volume of data, but only from a fraction of the overall Twitter message space. In fact, the Streaming API outperforms the old Search API methods in terms of message volume. Approximately 110,000 messages on Earth Hour were captured during the nationwide observation of Earth Hour starting at 8.30pm AEDST from the Australian East Coast, ending at 9.30pm AWDST in Perth[3]. In 2012, 220,000 messages were captured, about twice the 2011 amount.

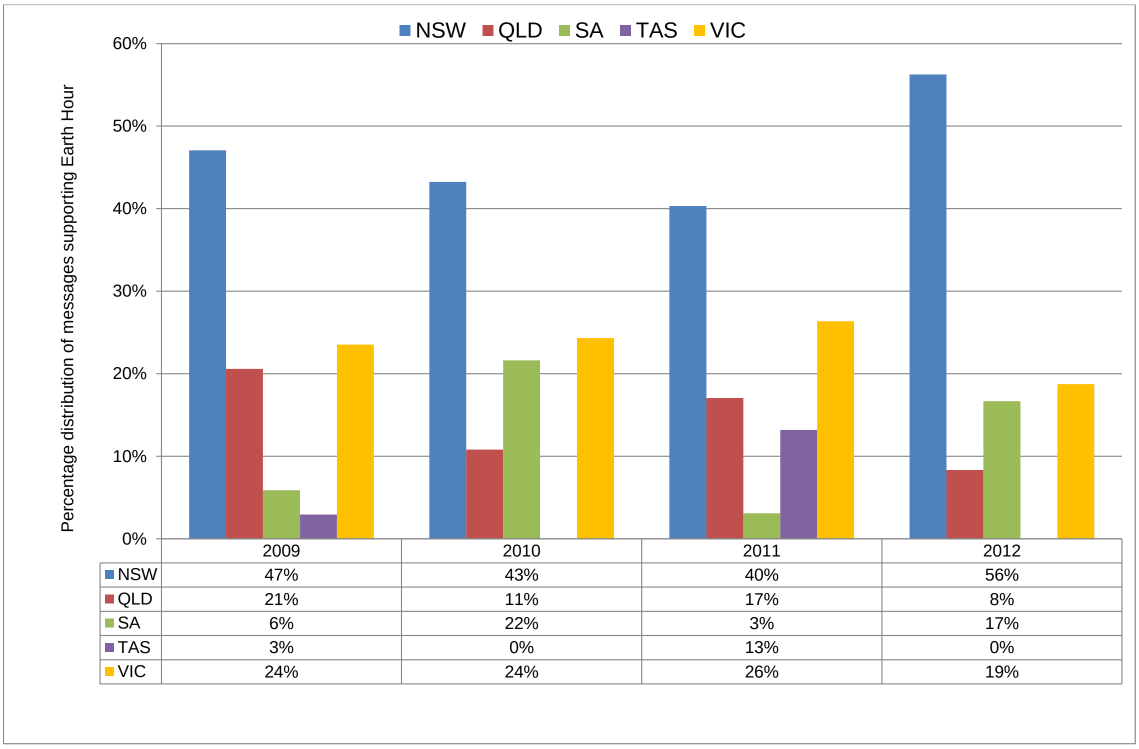
[2] Note on data collection: During the 2012 Earth Hour data collection, I was unable to collect data for a period of approximately nine minutes in the second half-hourly period for Queensland due to network connectivity issues. There were four messages from Queensland according to existing data; the figure was extrapolated to account for the loss of nine minutes of data.

[3] AEDST: Australian Eastern Daylight Saving Time; AEDST: Australian Western Daylight Saving Time.

2. **Shift in participation and hashtagging:** My methodology for detecting state-by-state activity is based on the original 2009 Earth Hour campaign. Originally, Twitter users were encouraged to promote the Earth Hour campaign by combining Earth Hour hashtags with their current location as a hashtag [William-Ross, 2009]. However, this self-organizing behavior has changed from 2010 onwards; since then, people rarely use the original hashtag convention to indicate locations. To keep in line with the existing methodology without introducing drastic changes, I ignored the hash character when detecting location names. Other experimental parameters were kept the same as my 2009 original study [Cheong and Lee, 2010d]. The list of place names as per Table 7.1, first published in [Cheong and Lee, 2010d], was kept constant.

3. **Noise:** Also, in the 2011 data, there was a noticeable increase in the number of tweets, mainly from New South Wales. Noise was present in the content of the tweets as 2011 Earth Hour coincided with the electoral defeat of Kristina Keneally in the 2011 New South Wales state elections.

As an aside, another way of visualizing the distribution of messages is by instead expressing the per-state message count as a percentage of the total messages for a given year. This is a form of normalization by accounting for the different sample sizes of Twitter messages throughout the four years (Figure 7.2).



	2009	2010	2011	2012
🟦 NSW	47%	43%	40%	56%
🟥 QLD	21%	11%	17%	8%
🟩 SA	6%	22%	3%	17%
🟪 TAS	3%	0%	13%	0%
🟨 VIC	24%	24%	26%	19%

Figure 7.2: Normalized percentage of Earth Hour tweets recorded per state, 2009–2012.

Experiment 7.3. *Determining correlation between per-state energy savings with respect to online Twitter activity (Study Goal 3).*

METHOD: To achieve this goal, I first obtain the nett reduction of energy consumption for each state (expressed as a percentage value over the normal average), for each of the Earth Hour years 2009–2012.

In a given year, the nett reduction for each state is matched with the aggregated Twitter usage count for that state. This is done across all four Earth Hours surveyed.

To check for statistical significance for a given years' data, I first calculate the Pearson product-moment correlation coefficient [Pearson, 1895], r, which was earlier defined in Equation 5.2. To recap: $r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$; in this experiment n is the number of states in this experiment; X_i and Y_i the Twitter message count and energy reduction percentage for the five states respectively; $\bar{X}$ the mean of the Twitter message count; s_X the standard deviation of the Twitter message count; $\bar{Y}$ the mean of the energy savings; and s_Y the standard deviation of the energy savings.

I derive the Student's t-value for a given year directly from the obtained Pearson's r value [Student, 1908; Rahman, 1968], where:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{7.2}$$

given n = number of observations (states). This is to test for statistical significance: i.e. to refute the null hypothesis that *there is zero correlation between the two samples (energy reduction versus tweet count) due to the small number of states (five) for which I have data for.*

RESULTS AND DISCUSSION: For each of the four years (2009–2012), the energy reduction for each state is listed along the Earth Hour tweet count in Table 7.6.

Table 7.6: Percent reduction of energy use during Earth Hours 2009–2012, and count of Twitter messages observed.

2009					
State	**NSW**	**QLD**	**SA**	**TAS**	**VIC**
Energy savings (%)	10.09	8.08	7.77	8.48	6.91
Twitter messages	48	21	6	3	24

2010					
State	**NSW**	**QLD**	**SA**	**TAS**	**VIC**
Energy savings (%)	4.74	6.76	6.76	-2.07	7.15
Twitter messages	16	4	8	0	9

2011					
State	**NSW**	**QLD**	**SA**	**TAS**	**VIC**
Energy savings (%)	8.70	7.48	15.07	7.97	7.04
Twitter messages	52	22	4	17	34

2012					
State	**NSW**	**QLD**	**SA**	**TAS**	**VIC**
Energy savings (%)	8.77	7.17	9.17	9.81	10.41
Twitter messages	27	4	8	0	9

From statistical significance testing, I have obtained Pearson r values for each year's samples, and corresponding Student t values in Table 7.7. For three of the years surveyed (2009–2011), the Pearson coefficient r has a value greater than 0.5. This suggests a correlation between the two parameters mentioned for the given data: i.e. the frequency of Twitter message activity is related to the percentage of energy savings in the Australian states.

Table 7.7: Results of statistical tests on energy savings (percentage) versus Twitter messages, for the 2009–2012 Earth Hours.

	Earth Hour			
Parameter	**2009**	**2010**	**2011**	**2012**
Pearson coefficient, r	0.5798	0.5225	0.5874	0.0513
t-value corresponding to r (d.o.f. = 3)	1.2325	1.0615	1.2570	0.0890
Significant at 5% level? (t-value $\geq$ 2.3534 required)	No	No	No	No
Significant at 10% level? (t-value $\geq$ 1.6377 required)	No	No	No	No
Significant at 15% level? (t-value $\geq$ 1.2498 required)	No	No	Yes	No
Significant at 20% level? (t-value $\geq$ 0.9784 required)	Yes	Yes	Yes	No

From the statistical point of view, the t-values obtained from the 2009–2012 data set is considered small enough to be statistically significant at the 5% and 10% levels. This is because all four sets (2009–2012) have t-values less than 2.3534 (needed for 5% level of significance at 3 d.o.f.) and 1.6377 (needed for 10% level of significance, at 3 d.o.f.).

However, the correlation between the 2011 Earth Hour tweets and energy drop is statistically significant at the 15% level (3 d.o.f.). The correlations for the 2009 and 2010 Earth Hours are significant at the 20% level (again, both with 3 d.o.f.).

A scatter-plot of the data from Table 7.6 visually illustrates the two variables in the current experiment over the period of 2009–2012 in Figure 7.3.

Overall, the results from Experiment 7.3 suggest a correlation ($r^2 > 0.5$) between the *Twitter participation of users* (number of Earth Hour tweets per state) and the energy savings recorded per state during Earth Hour for each of the years 2009–2012. However, the correlations are statistically significant only at lower confidence levels. A major factor contributing to this is the small number of states in which I have data for, which in turn, reduces the degrees of freedom in determining statistical significance.

Experiment 7.4. *Discovering the relationship between per-state Twitter activity with state population (Goal 4).*

METHOD: To complete the final goal of this longitudinal study, I compare the *Twitter participation of users* (i.e. per-state tally of Earth Hour tweets) to the population of each of the specified Australian states. This ratio provides an insight into the adoption rate of microblogging (expressed by the proportion of people participating in the Earth Hour Twitter campaign) relative to the size of each state in terms of its population.

The population data on a state-by-state basis were obtained from the Australian Bureau of Statistics [Australian Bureau of Statistics, 2009] as it is an authoritative source

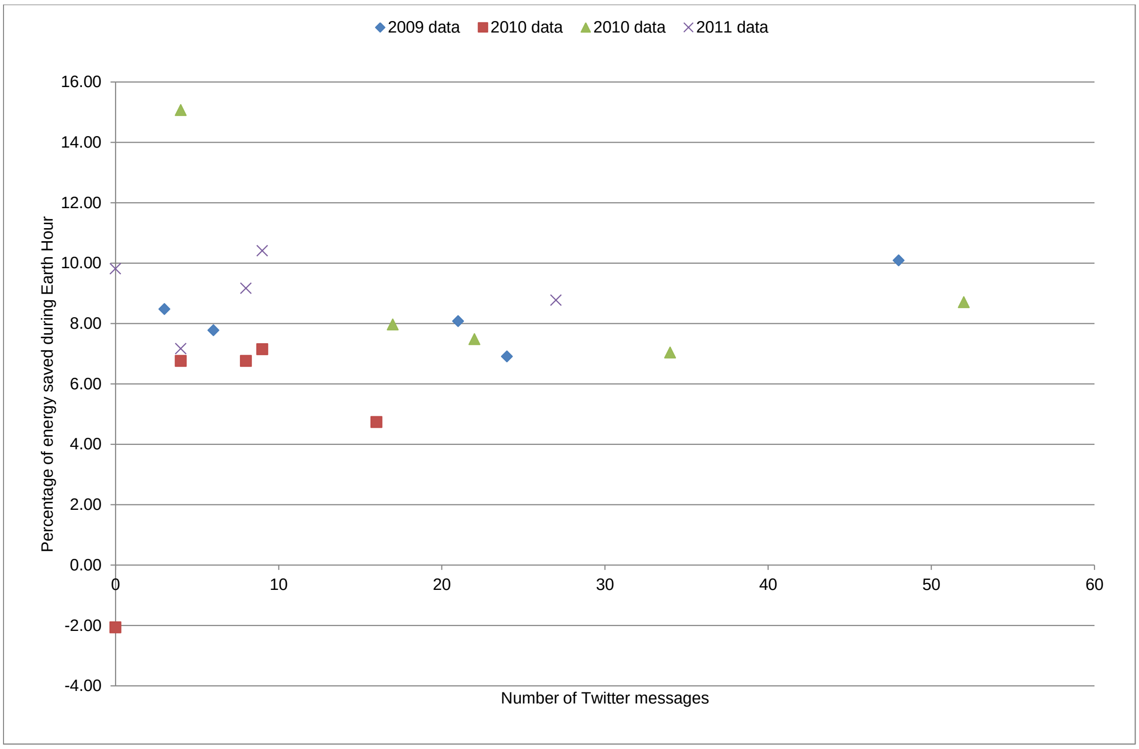


Figure 7.3: Scatter-plot of energy savings (expressed as percentages) versus count of Twitter messages for the Australian states.

of population and demographic statistics. The version of population statistics used for a given year is current as of Earth Hour for that year.

The calculation of the Pearson product-moment correlation coefficient (r as defined by Equation 5.2) for each years' worth of paired tweet-versus-population data is performed as per Section 7.3, to determine the degree of correlation between the two variables. Again, to check for statistical significance due to the small number of Australian states sampled, I calculate Student's t-value (Equation 7.2) given the r coefficients for each year. This is to test for statistical significance and refute the null hypothesis of *zero correlation between per-state tweet count and population size.*

Results and Discussion: The yearly population data are presented as a comparison with Earth Hour Twitter message count in Table 7.8. The same data are visualized as a scatter-plot in Figure 7.4.

The results of significance testing for all four Earth Hours are as per Table 7.9. It is shown that the t-values — 5.6483 (for 2009) and 3.1995 (for 2011) — corresponding to the 2009 and 2011 Earth Hour's Pearson coefficients (0.9561 and 0.8794 respectively) is higher compared to an expected t-value of 2.3534 (5% level of significance, at 3 degrees of freedom); indicating that the relationship between the two variables in the 2009 Earth Hour is statistically significant.

The 2010 and 2012 data sets have t-values of 2.3001 and 2.1379 respectively, which are slightly short of the target value of 2.3534, which is small to be considered statistically

Table 7.8: Year-by-year comparison of Australian state population (millions) with total Earth Hour Twitter messages for a given state.

2009					
State	**NSW**	**QLD**	**SA**	**TAS**	**VIC**
Population (millions)	7.0414	4.3495	1.612	0.5003	5.3648
Twitter messages	48	21	6	3	24

2010					
State	**NSW**	**QLD**	**SA**	**TAS**	**VIC**
Population (millions)	7.1915	4.473	1.6339	0.5054	5.4964
Twitter messages	16	4	8	0	9

2011					
State	**NSW**	**QLD**	**SA**	**TAS**	**VIC**
Population (millions)	7.2354	4.5323	1.6478	0.5085	5.5671
Twitter messages	52	22	4	17	34

2012					
State	**NSW**	**QLD**	**SA**	**TAS**	**VIC**
Population (millions)	7.3175	4.5994	1.6598	0.511	5.6409
Twitter messages	27	4	8	0	9

Table 7.9: Results of statistical tests on Twitter messages versus population data, for the 2009–2012 Earth Hours

	Earth Hour			
Parameter	**2009**	**2010**	**2011**	**2012**
Pearson coefficient, r	0.9561	0.7988	0.8794	0.7770
t-value corresponding to r (d.o.f. = 3)	5.6483	2.3001	3.1995	2.1379
Significant at 5% level? (t-value $\geq$ 2.3534 required)	Yes	No	Yes	No
Significant at 10% level? (t-value $\geq$ 1.6377 required)	Yes	Yes	Yes	Yes

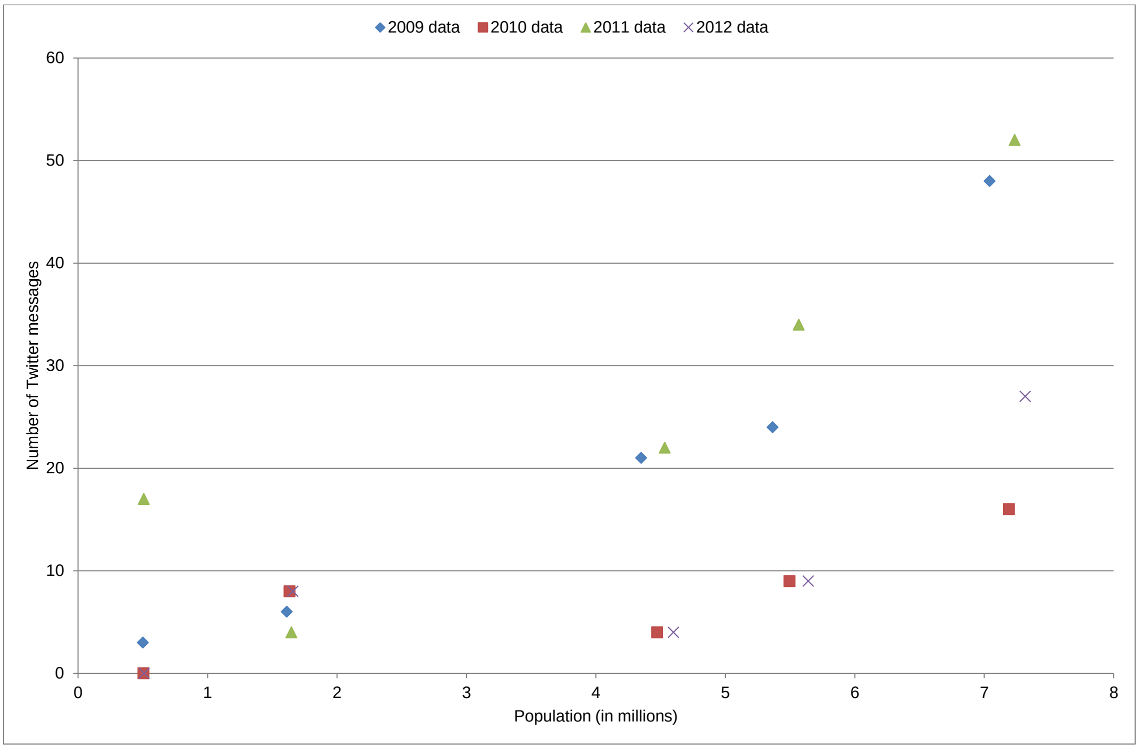


Figure 7.4: Scatter-plot of state population (millions) and the Twitter messages for the Australian states.

significant at $\alpha = 0.05$. However, their values of t are statistically significant at a 10% level, $\alpha = 0.1$, as they exceed the required $t = 1.6377$ for 10% significance.

One way to interpret a potential correlation between the number of Earth Hour tweets per state versus that state's population is that the usage rate of Twitter and such microblogging technologies depend on how populated a particular locale is.

The high significance of the 2009 and 2011 data sets illustrate this well. However, the 2010 and 2012 data set is not statistically significant enough; my reasoning is that the lack of Twitter message samples and significant disengagement with Earth Hour participation via Twitter caused this to happen. Nevertheless, such a metric has potential as a basis for work on measuring the penetration rate of social media, microblogging, and related technologies in Australia as well as other geographic regions.

7.1.7 Discussion

A few limitations have been identified in this longitudinal study. First of all, there is a dearth of Twitter messages regarding Earth Hour especially in the 2010 and 2012 editions. A case in point would be the state of Tasmania having zero messages in 2010 (as opposed to three for the 2009 Earth Hour). The limitation of not having a complete data set of all messages heavily restricts the number of tweet samples that can be obtained. This limitation is hard to solve, but if the complete set of such messages was made available by Twitter Inc., I would have more accurate data on which to base the analysis.

The limited number of Australian states for which I have data readily available (as there are only five data points) is another issue that could be improved in future work.

Future work could also expand the scope of this study to include a country-by-country comparison, or region-by-region comparison, for example.

Another study limitation is that for the years 2010–2012, I use my original 2009 case study methodology, published in Cheong and Lee [2010d]. In 2009, only the tweets with both the `#earthhour` and `#location` hashtags were harvested for consistency. However, since 2010, I have observed that only a few people use the same originally-proposed hashtag notation — which was proposed during the 2009 Earth Hour [William-Ross, 2009] — to tag their tweets. The rest of the users provide abbreviations of location names, forget to designate the location with a hashtag, or depend on the functionality of their mobile Twitter clients to publish geographic coordinates along with their tweets.

For this particular issue, suggestions of improvements to in future work include complementing the hashtag notation of place names with location metadata, such as via my *Two-phase Hybrid Geocoding* [Cheong et al., 2012b] approach (Section 4.6.2). Coordinates or place names found in user metadata can be geocoded (Algorithm 4.2); in the absence of which, locations can be inferred based on a tweet author's free-form location text string with the help of Algorithm 4.3. Alternatively, message metadata can be used to provide an educated guess of the user's location, such as user language, locale, and time zone information: the latter has been used with varying degrees of success in current research [Krishnamurthy et al., 2008; Schafer, 2010; Kwak et al., 2010; Java et al., 2009].

7.1.8 Study Conclusion and Future Work

This case study was built upon my original study conducted on the 2009 Earth Hour, published as [Cheong and Lee, 2010d]. I was able to convert my original study [Cheong and Lee, 2010d] into a longitudinal one by augmenting it with data collected from the 2010–2012 Earth Hour campaigns. Several changes have been introduced, such as the addition of statistical significance tests and an improved data collecting framework.

The longitudinal study, as seen in this entire section, consisted of four goals. The following list summarizes the post-experimental conclusions with respect to each of the four study goals:

- *Study Goal 1 conclusion*: Analysis of real-world phenomena can be done efficiently by utilizing public records; in this case, energy market data was obtained to determine any difference in energy consumption during the 2009–2012 Earth Hours.

- *Study Goal 2 conclusion*: The drop of energy consumption during the 2009–2012 Earth Hours are statistically significant (at a 95% confidence level). This indicates a successful strategy by the Earth Hour organizers in promoting energy conservation, generating 'buzz' among the populace and commitments from local governments worldwide. A limitation exists in that the experiment is only confined to five major Australian states for which I have energy market data.

- *Study Goal 3 conclusion*: There is a possible link between Australian Twitter usage patterns and the efficacy of a Twitter-based (and social media-based) global activism campaign that is Earth Hour. Due to study limitations, however, the obtained results are too small to be statistically significant.

- *Study Goal 4 conclusion*: The claim that Twitter activity is a good reflector of the real-world population was found to be statistically significant (at least 90% confidence level for all cases).

In closing, Twitter activity can indeed be translated into action in a real-world social system. Likewise, human behavior in the real-world especially during such campaigns can be seen manifested online in terms of microblog chatter. Future work related to this study include studying how behavior of other online social systems can be mapped to the real world; how microblogging can be an avenue for self-expression; measuring real-life sentiment and gauging response via microblog posts; and exploring other avenues of mass coordination via microblogging and social network technologies.

7.2 Twitter in Terrorism Informatics & Public Response

From the prior case study of real-life activism campaign via online interactions on Twitter, this next case study, first published as [Cheong and Lee, 2011], tackles a similar but far yet sinister problem in modern times. This section will highlight on manifestations of civilian response to terrorism events via Twitter, and its consequences. I propose a potential framework based on the Twitter API and adapting the user and message inferences covered in Chapter 4 for *terrorism informatics* — i.e. the study of terrorism-related information, its collection, and management — on Twitter.

7.2.1 Background: Twitter and terrorism

Twitter has, of late, become a medium of information sharing and dissemination, and also an avenue to break news faster than traditional news outlets. It is interesting to note that Twitter is not only potentially beneficial in terrorism informatics and identifying threats [The Associated Press, 2009] but also has been identified as a potential facilitator for coordinating activities of terrorism [Musil, 2008] and a threat to security [Entous, 2009].

In terrorism informatics, tracking, location and time of activities vary significantly and thus become extremely hard to predict. Intelligent information sharing techniques, applied to unstructured content of texts, can lead to the discovery of hidden rare patterns for real-world disaster and crisis management situations.

7.2.2 Motivations

I posited earlier in Chapter 4 that tweets and their associated user and message metadata, though with noisy and unstructured content, can exhibit emergent characteristics with regards to the social network dynamic [Goolsby, 2009; Huberman et al., 2008a] and can be used to indicate sentiments/behavior of users discussing a particular topic [Cheong and Lee, 2009; Shamma et al., 2009].

Based on these observations, the goals of this study are:

- **Study Goal 1**: To study existing literature on terrorism informatics to see how Twitter, with its 140-character limit and not-readily-obvious metadata, can be effectively used in terrorism response informatics.

- **Study Goal 2**: To build an experimental framework based on metadata inference methods to track and summarize the reaction of the civilian population on Twitter in the aftermath of terrorist activity.

- **Study Goal 3**: To propose a number of visualization methods to graphically analyze the inferences generated from *Study Goal 2* above, and to integrate such methods into the experimental framework of *Study Goal 2*.

- **Study Goal 4**: Lastly, to conduct a simulation based on synthetic data exhibiting characteristics found in terrorism informatics literature. The framework (proposed in *Study Goal 2*) and visualization techniques (resulting from **Study Goal 3**) are tested on simulated Twitter activity; which was synthesized from real-world Twitter metadata using properties found in current literature on terrorism informatics.

Knowledge on terrorist activities can be extracted and integrated with existing data sources to provide authorities with a richer source of information to both chronicle current threats and learn more about them.

7.2.3 Literature Review: Twitter in Terrorism Informatics

There is limited research specific to the usage of Twitter in terrorism informatics. From the extant literature (Chapter 3), I have identified several academic studies regarding the properties of the Twitter user base and its goings-on that is relevant to terrorism informatics.

Analyses of properties of the Twitter user base [Java et al., 2009; Krishnamurthy et al., 2008; Huberman et al., 2008a] touch on aspects such as growth rate, geographic profile, user habits, and the social network of the Twitter community as a whole. From a humanities perspective, Mischaud [2007] and Erickson [2008] studied motivations behind information sharing on Twitter and concluded that Twitter is used for information sharing and broadcasting of everyday goings-on. This finding, in turn, leads me to postulate that Twitter can be studied to analyze the sentiment, current condition, and response of civilians affected by a sudden act of terrorism.

On a topic closely related to terrorism informatics, Hughes and Palen [2009] have surveyed the adoption and use of Twitter during mass convergence and emergency events, specifically those involving national security, in the perspective of *crisis informatics*. They allude that Twitter messages exhibit “features of information dissemination [supporting] information broadcasting and brokerage” [Hughes and Palen, 2009], and Twitter may be used as a tool for emergency response and communication by the authorities in order to provide aid and counter disinformation. Related research on the Canadian Red River Valley floods of 2009 Starbird et al. [2010] have detected patterns of social information and self-organization by users discussing the flood. Starbird et al. [2010] notice a pattern of

"...commentary and the sharing of higher-level information" and a combination of tweets with authoritative news sources in their research sample, solidifying the claim that Twitter can be used to get a feel for civilian response after an event of terror has occurred.

Jungherr [2009] detailed the role of Twitter in social activism and looked into case studies whereby Twitter was instrumental in disseminating information on terrorist attacks, political dissent, and acts of oppression. Goolsby [2009] has also stated that Twitter can be used "...[to] cover crucial events [in situations like] state terrorism". Their quote, "Mumbai has shown the potential for using microblogging systems like Twitter in breaking events..." [Goolsby, 2009] has aptly summarized the role of Twitter in such situations.

Based on the above prior work, I propose a framework to adopt Twitter to a study on the reactions, sentiments, and communication of civilians in response to terrorist attacks. I describe a novel application of Twitter metadata analysis (combining demographic analysis with elements of textual analysis on tweets), wherein I propose a four-phase framework for using Twitter to visualize civilian response to a terror attack. The resulting extracted information can then be used by the authorities for the purposes of rapid detection, response, and recovery as posited by Tien's decision informatics paradigm [Tien, 2005].

7.2.4 Empirical Observations on Twitter during Terror Events

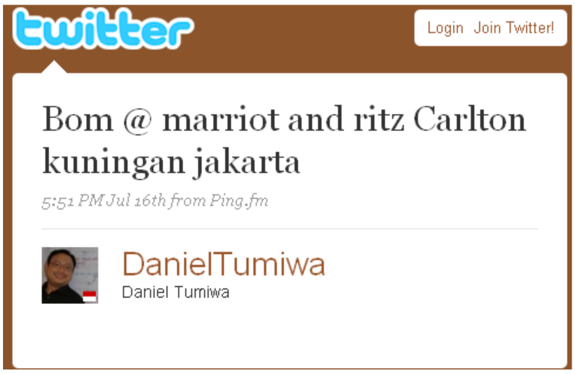
In preparing my framework for this study, I draw upon existing empirical findings seen in civilian response to recent terrorism activity, with a main focus on urban terrorism as it "...[produces] the most visible impact" [Tien, 2005]. I also capitalize upon the existing trend of information needs and sharing via microblogging and online social networks. The motivation and foundation of my framework is as follows.

Twitter has been known to be one of the channels where civilians break news of terrorist activities and use it as a method to notify the public of any latest updates, cries for help, and as an information source for the authorities. Such information can come in the form of a plain text tweet or even related content or media, for example, photos and video. Oftentimes, the peak of activity related to the sudden spike of a breaking news story can cause it to be promoted into the Twitter's Trending topics list [Cheong and Lee, 2009; Cheong, 2009].

Examples would be:

- **The 2008 Mumbai attacks**: News of the attacks were first reported by citizen journalists on location via Twitter [Beaumont, 2008; Goolsby, 2009].
- **The Jakarta bombings of July 2009**: Twitter was the first medium that broke the news of the incident [Cashmore, 2009a; Saputra and Leitsinger, 2009]. The first few images of the tragedy were broadcast to the general public via a user posting on *TwitPic* (Figure 7.5).

Similar examples also show civilian reporting of accidents, crime, and other forms of disaster via Twitter and other Web 2.0 social networks, news aggregators, and media-sharing services. It is useful to gain an insight into such cases where terrorist activity



Bom @ marriot and ritz Carlton kuningan jakarta

5:51 PM Jul 16th from Ping.fm

DanielTumiwa

Daniel Tumiwa

Figure 7.5: (left) Twitter post by user `@DanielTumiwa` <`http://twitter.com/DanielTumiwa/status/2679572777`>; (right) *TwitPic* photo by user `@GaluhRiyadi` <`http://twitpic.com/alt25`>) chronicling the Jakarta bombings.

does not impact the public directly; but rather in the form of collateral damage, which includes:

- **Hudson River plane crash:** User `@jkrums` sent a tweet containing an image link to *TwitPic* to deliver the first few glimpses of the tragedy to the outside world [Terdiman, 2009].

- **Assassination of Neda:** The breaking of the news of the assassination of Neda, an Iraqi civilian, in response to the crackdown of the aftermath of the 2009 Iranian Election protests [Fleishman, 2009]. Immediately after the event, news spread through Twitter and other Web 2.0 channels. YouTube links were found to contain clips of the assassination and passing away of Neda, as opposed to the Iranian mainstream media's lack of coverage due to a severely-restricted press environment.

I propose a framework for information extraction from Twitter messages for terrorism informatics, consisting of four distinct phases. This is built up on prior work so far in Chapter 4 of this thesis, specifically the usage of a data gathering framework using Twitter's on-demand `search` and REST-`user` APIs (Section 4.1), and inference algorithms (Sections 4.6, 4.7, and 4.8).

7.2.5 The Proposed Terrorism Informatics Framework

In this section, I describe the implementation of my proposed framework. A prototype implementation of this framework was created in the Perl programming language, using the `Net::Twitter` wrapper for the Twitter API [Cheong and Lee, 2010c].

A high-level overview of the framework is illustrated in Figure 7.6. Briefly, the various phases are (from top-to-bottom in Figure 7.6:

- **Phase 1**: This phase involves firstly identifying terrorism-related trending topics from Twitter, via the Trending Topics API.

- **Phase 2**: Given a list of terrorism-related topics as a result of *Phase 1*, the tweets matching such topics in *Phase 2* are continuously fetched. This is done by virtue of either the Streaming API which is preferable as of late 2010 (see Section 4.2.3), or the on-demand `search` and REST-`user` APIs in which my study was based on when it was first published.

- **Phase 3**: In this phase, post-processing is performed on the message metadata found in *Phase 2*. Any missing user metadata is obtained from Twitter (via the REST-`user` API), particularly if the older `search` API was used. Then, inference and sentiment detection algorithms specifically for terrorism informatics, are run against both the user and message metadata.

- **Phase 4**: The final phase consists of visualization and clustering of the resulting inferences and sentiments obtained in *Phase 3*. Such visualization/clustering techniques allow for easy understanding and interpretation of the wealth of information that can be obtained from this framework.

In the following sections, I dissect my proposed framework into the four individual phases, and describe each phase in detail.

7.2.6 *Framework Phase 1*: Breaking news

In this phase, topics and hashtags discussed on Twitter are analyzed by querying the Twitter Trending Topics list, a list of very frequently discussed topics updated on a regular basis. By monitoring the most talked about messages at any given time for signs of potential terrorist activity, one can use Twitter to chronicle the civilian response to such a threat from the moment news first breaks out. Figure 7.7 illustrates *Phase 1* of the framework.

Twitter has an API that allows tracking of the ten most talked about, or trending topics. By analyzing this list of topics for breaking news stories regarding terrorist activity, potential mentions of civilian reaction towards terrorist activity can be identified using this list. Hence, processing can be narrowed down to those specific messages, as seen in the cases of the Mumbai [Beaumont, 2008] and Jakarta [Saputra and Leitsinger, 2009; Cashmore, 2009a] bombings, where the keywords `Mumbai` and `Jakarta` quickly broke ranks to become one of the top trending topics.

In this initial phase, I propose the querying of the Twitter `trends` API at a predetermined interval (for example every ten minutes). By scanning through the topics list for names of places and identifying trends which discuss about a flurry of activity at any single place (e.g. names of towns, cities), as in the case of the Mumbai and Jakarta bombings, I can isolate them as potential places where a terrorism attempt might have been executed. This is based on Guy et al. [2010], who proposed the notion that a "geographically concentrated spike of tweets" could draw focus on a certain location, indicating something major is happening there at a given point in time.

To automatically analyze the names in trending topics, I use location finding methods, such as the originally proposed Google Geocoder API in [Cheong and Lee, 2010c], or my proposed *Two-phase Hybrid Geolocation* approach (Section 4.6.2) running on a scalable

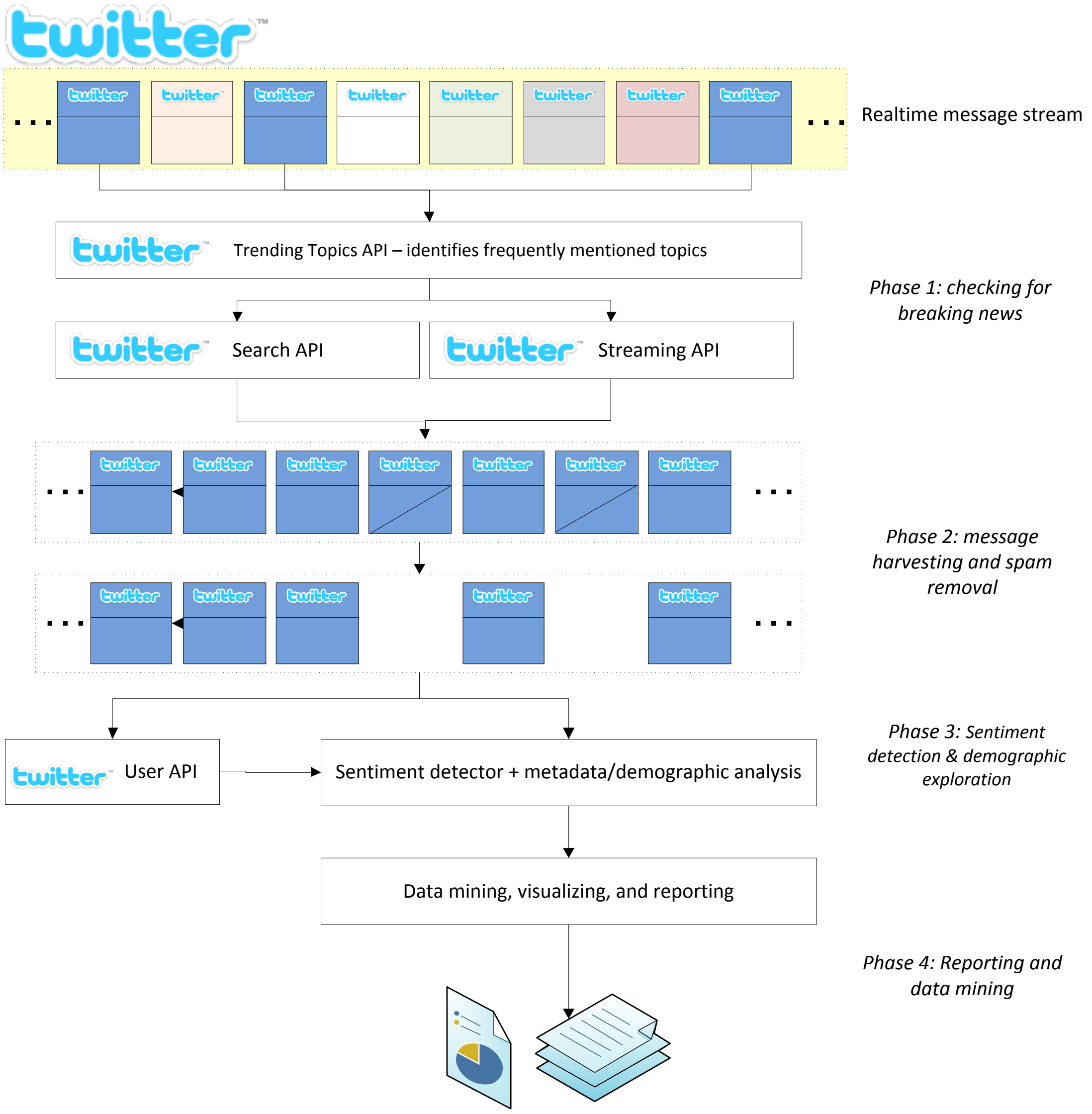

twitter
Realtime message stream
Trending Topics API – identifies frequently mentioned topics
Phase 1: checking for breaking news
Search API
Streaming API
Phase 2: message harvesting and spam removal
User API
Sentiment detector + metadata/demographic analysis
Phase 3: Sentiment detection & demographic exploration
Data mining, visualizing, and reporting
Phase 4: Reporting and data mining

Figure 7.6: High-level overview of my proposed terrorism informatics framework.

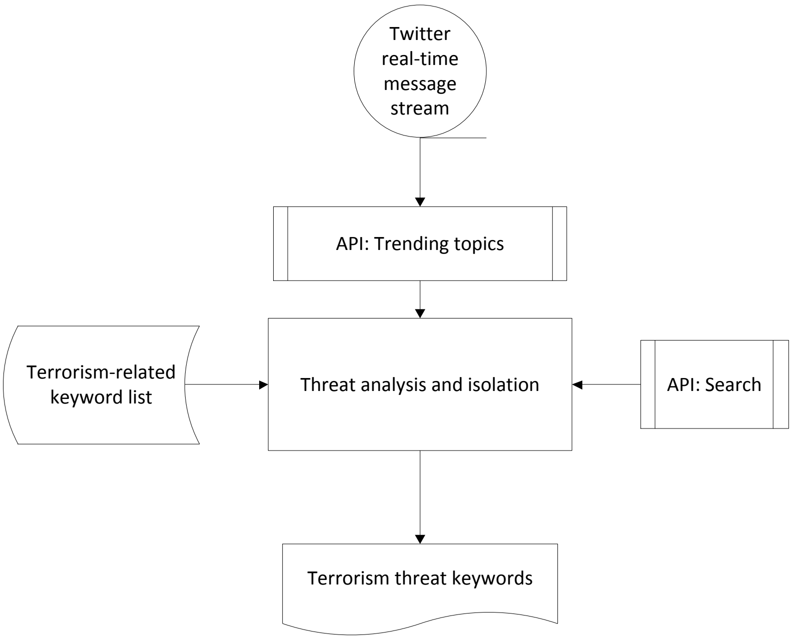

Twitter real-time message stream
API: Trending topics
Terrorism-related keyword list
Threat analysis and isolation
API: Search
Terrorism threat keywords

Figure 7.7: Framework diagram for *Phase 1*, illustrating the processing of breaking news.

cloud computing platform to detect mentions of geographic locations, which is also a similar approach by Guy et al. [2010].

The potential problem of disambiguating between proper names: e.g. between people, cities/location names could arise judging from the fact that certain human names (e.g. *Victoria*) double as a location name. Potential workarounds to this expected issue is performing a name look-up with a simple frequency-based gender detection algorithm [Cheong and Lee, 2010c; Cheong et al., 2012b] (hitherto discussed in Section 4.6.1). If the frequency of use as a proper male/female name falls below a particular threshold, geocoding can be applied on it (cf. Section 4.6.2) to assert that it is indeed a geographic location.

Once a particular location has been narrowed down, recent tweets that contain the location name in the message content are retrieved. Then, the corpus of retrieved posts is scanned for terrorism-related keywords (covered in *Phase 3*, Section 7.2.8) to positively identify a threat, as can be seen in case studies of terrorist activity in Mumbai and Jakarta.

To improve the findings from this phase, I also include the monitoring of trending topics with a list of keywords frequently mentioned during potential terror attacks (to be discussed in greater detail in *Phase 3*). In my opinion, the mentioning of such keywords which are uncommon in everyday topics [Cheong, 2009] but prevalent in terror attacks [The Sunshine Press, 2009] could potentially reveal that a terror attack is occurring, even without explicitly stating the exact location. This is similar to the use of a baseline frequency for geographically-based tweets to measure anomalies and detect breaking events [Abrol and Khan, 2010].

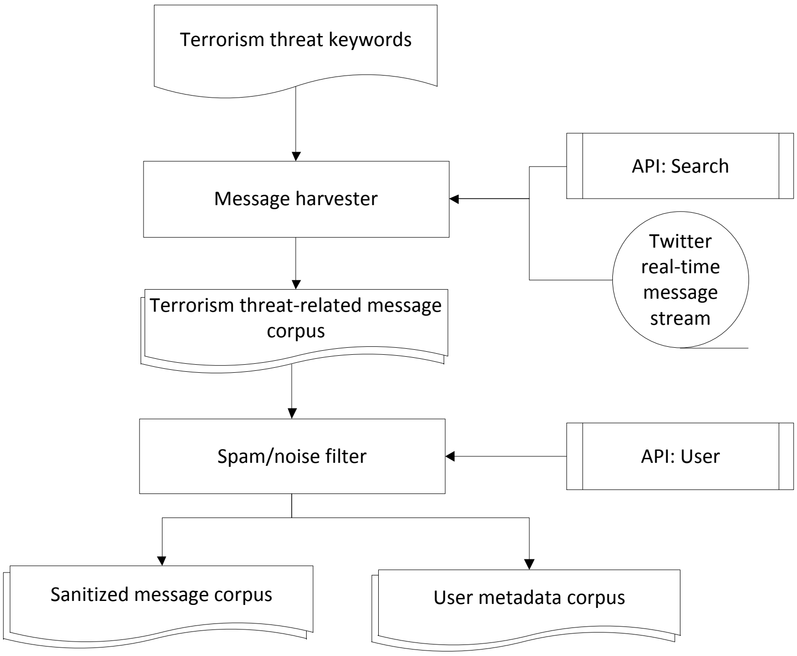


Figure 7.8: Framework diagram for *Phase 2*, illustrating data harvesting and the spam filtering process.

7.2.7 *Framework Phase 2*: Data harvesting and spam filtering

Once the location for the threat has been identified, message harvesting is performed on all related Twitter messages. Figure 7.8 illustrates this process. For the harvesting, Twitter's `search` API or alternatively the Streaming API can be used as data sources. The `search` API is used to query the topic and its past discussion right up to the terrorism event; on the other hand the Streaming API is used for monitor the real-time chatter on Twitter to capture tweets about the event as it happens.

Related literature has already determined that spam or unrelated noise [Cheong and Lee, 2009, 2010c; Krishnamurthy et al., 2008; Thomas et al., 2011; Metaxas and Mustafaraj, 2010] and real-world case studies [Cashmore, 2009b; Relax News, 2009] are commonplace in Twitter and can thus pollute the content stream. An example would be keyword injection by bot programs as part of spamming activity [Cheong and Lee, 2010b; Lee et al., 2010; Metaxas and Mustafaraj, 2010]. A method to dispose of such messages is thus required.

Characteristics of spammer users on Twitter have been identified in prior research (Chapter 3.3.4). Certain emergent properties and usage characteristics exhibited by certain classes of Twitter messages/users indicate that the user is likely to be contributing to noise or spam in the information stream. Examples of this include relative newness of a Twitter account, low degree of profile customizations, and omission of certain biographic data in the Twitter user profile [Cheong and Lee, 2010c; Thomas et al., 2011; Barracuda Networks, Inc., 2010; Metaxas and Mustafaraj, 2010; Collins, 2009; Lee et al., 2010]. Based on the above knowledge, I propose a novel noise-reduction filter to discard messages suspected of polluting the message stream with noise or spam.

First of all, the user information for each author of a terror-response tweet needs to be obtained. As of time of writing this thesis, Twitter's `search` API has embedded message metadata, similar to the Streaming API, eliminating the need for a REST-`user` lookup. However, if legacy data was used, e.g. historical tweets from around the time this study was published [Cheong and Lee, 2011], user metadata for a given tweet needs to be separately accessed using Twitter's REST-`user` API. (Any user data obtained this way is cached to reduce API calls for the same user in the future without counting towards a rate limit, cf. Section 4.1.3). As for metadata harvested from the Streaming API, both user and message metadata are provided, thereby eliminating the need for an extra user lookup.

Based on prior work [Cheong and Lee, 2010c; Dearman et al., 2008; Cheong et al., 2012b], I identify the *device classes* from `source` strings used to post Twitter messages that are least likely to contain spam. Examples would be the web interface, mobile devices, and social media programs. This is in contrast with RSS feeds and other Twitter content generators which are highly likely to contain spam and contribute to overall noise in the Twitter feed; such generated tweets would be removed from the message corpus.

My proposed noise-reduction filter will remove messages by users who have been on Twitter for less than a particular timeframe. For the purposes of my experimental simulation (to be discussed in *Phase 4* later in Section 7.2.9), I use the time-frame of a week to filter out newly-created bot accounts, suspected of automatically generating spam on Twitter. By directly excluding the content generated by such users, and prioritizing the content created by *de facto*/legitimate users, the percentage of spam is greatly reduced.

Other possible noise-reduction techniques can include other metrics — such as profile customization scores, follower/friend ratios, and user activity ratios (Sections 4.7.1, 4.7.3, and 4.7.4 respectively) — for the detection of anomalous users. The sanitized pool of messages and the user metadata is then saved to disk for further processing.

7.2.8 *Framework Phase 3*: Sentiment detection and demographic exploration of the message pool

In this phase, I propose a two-step approach to exploring the latent information in the sanitized message pool, as illustrated in Figure 7.9.

The first would be performing sentiment analysis, where the reaction of the general civilian population would be monitored. Sentiment detection methods have been successfully applied to Twitter in prior work. Examples would be to gauge the public sentiment in politics [Shamma et al., 2009] and opinion in marketing [Jansen et al., 2009a].

For the purposes of this terrorism informatics study, I devise a simple new sentiment detection mechanism. My proposed terrorism informatics sentiment detector detects common keywords related to public reaction and descriptions of terrorism. This is done by feeding a list of potential sentiment-related keywords which will then tag the messages based on category of the keywords detected in the incoming message (Figure 7.9).

There are no readily available list of such words, especially when used in the context of day-to-day communication. Hence, there is a need for the construction of a usable word

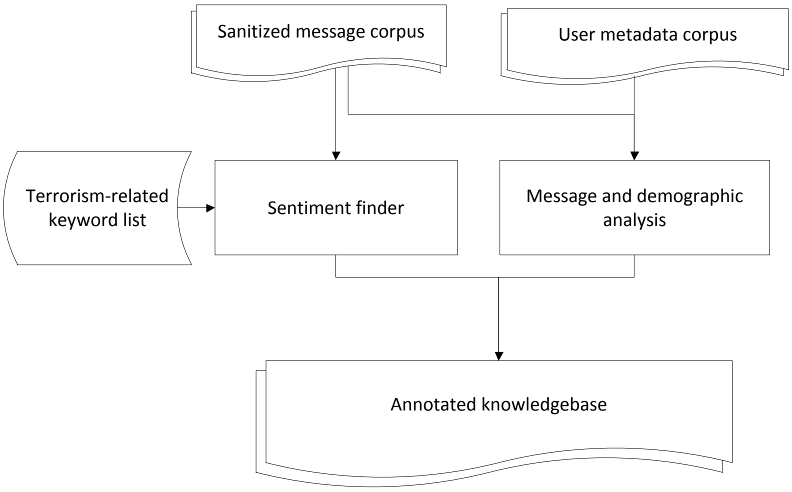


Figure 7.9: Framework diagram illustrating sentiment detection and demographic exploration in *Phase 3*.

list which contains sentiment keywords on Twitter during potential terror events, which I have performed and documented in Experiment 7.5.

Experiment 7.5. *Building a list of keywords that indicate sentiment on Twitter during terrorism events, organized by category.*

Method: I studied existing literature pertaining to terrorism informatics and real-world communications data captured during terror events to determine keywords (and their corresponding categories). The following three steps were taken in the construction of this list of sentiment keywords:

1. Studying the common responses evoked in civilian survivors, first-responders, and people affected by the aftermath of terror activities, as documented by Beutler et al. [2006].

2. To provide a more systematic data set and to capture the real-world communication patterns by survivors and observers in a real-world terrorism scenario, I draw upon the research by Clark [2009] on a corpus of 448,358 pager messages [The Sunshine Press, 2009] captured during the 9/11 terror attacks; where the most 100 frequently occurring key phrases in this data set were ranked. From the list of key phrases found in Clark [2009], I added related frequently-used keywords (from the original 448,358 message dump) to produce a set of systematic root words/phrases for capturing terrorism-related sentiments.

3. Additional synonyms pertaining to words and phrases found in steps 1 and 2 above were obtained using the WordNet lexical analysis tool [Miller et al., 1990].

Results and Discussion: The proposed categories, together with some example keywords, are listed in Table 7.10.

Table 7.10: List of categories for sentiment analysis. (Keywords marked with * are common internet abbreviations related to the other keywords in the same category).

Category	**Keywords**
Emotion: fear/anxiety	*anxiety/anxious, catastrophic, concern, disaster, emergency, fear, insecure, panic, scared, terror, threat, trouble, warning, worry*
Emotion: shock	*(taken) aback, floor, god bless, omg*, shock, stun, sudden, wtf*, wth**
Response	*act, asap*, escape, evacuate, flee, help, hide, run*
Need for information and updates	*breaking news, call, foul play, incident, phone, report, situation, unconfirmed*
Assessment: threats	*accident, attack, bomb, bullet, collapse, crash, explode/explosion, fire, gun, hijack, hit, hostage, plane, responsibility/responsible, rifle, shot/shoot, struck, suicide, terrorist*
Assessment: casualties	*blood, body/bodies, corpses, dead, casualties, injury/injure, kill, wounded*
Response and law enforcement	*action, ambulance, command, medic, operation, planes, police/cops/FBI/security, recover, rescue, response, restore, safe, safety, save, shut, stay, survive, suspend*

This data set, in my opinion, is a voluminous set of recent terror-related key phrases used by civilians in communication during/after a terrorist attack that describes their reaction and sentiments, justifying the use of Clark's research [Clark, 2009].

However, it is pertinent to note that the list of words here is by no means exhaustive; prior research found that the difficulties of predicting the possible range of victims' responses to terrorism is due to:

> "*...the unavailability of systematic, empirical research on the events that immediately follow a terrorist attack [due to the fact that] these attacks are infrequent and unexpected.*" [Beutler et al., 2006], based on [Neria et al., 2004].

The second step in this phase is to extract, filter and process metadata attached to the Twitter messages sent out. As identified earlier, several attributes can be used to identify physical properties of the authors behind tweets, as well as cull extra information [Cheong and Lee, 2010c; Hughes and Palen, 2009; Java et al., 2009; Krishnamurthy et al., 2008] related to the event that can be used to improve the assessment of the terrorism event and also to assist in immediate decision-support by the appropriate authorities.

In my proposed framework, derived attributes and properties acquired from the sanitized corpus are divided into several categories, as listed in Table 7.11.

Once such attributes have been identified, they will be annotated (as with the sentiment analyses above). The resulting annotations from sentimental analysis and user metadata will be combined with the original message corpus and stored in a knowledge base ready for further reporting, visualization, or pattern recognition in *Phase 4*.

Table 7.11: List of properties and derived attributes from the available user metadata.

Category	**Attributes sought**
Spatiotemporal properties	**Time and date of Twitter message by a user:** Because the time and date is stamped by the Twitter service consistently using GMT as its reference offset, the accuracy of individual computer/device clocks do not matter. **Location information:** Messages can be used as real-time source of geographic information in tracking the aftermath of terrorist threat if location information about the user corresponds to the location of terrorist activity or collateral damage. The most credible messages among these are geo-tagged with latitude/longitude information published by Twitter clients which support GPS technology, for example mobile devices. These can then be parsed directly by e.g. using the GeoRSS specification [Open Geospatial Consortium Inc., 2006]. With the advancement of the Streaming API and Places API, newer tweets can contain embedded per-message location data. The spatiotemporal properties mentioned above enable pinpointing of first-responders and civilians immediately after a terror event has occurred.
Gender	**Gender of a tweet's author:** This can be predicted by running the name provided on his/her Twitter user profile [Cheong and Lee, 2010c] through a frequency-based ranking algorithm. This is useful in identifying the general demographic of civilians affected by a terrorism scenario.
User mobility state	**Device class of the Twitter client used:** Based on the source metadata item of a tweet, the device class is ascertained to determine whether the user was mobile or in a fixed location [Dearman et al., 2008]. This allows one to determine the situation of the message author (e.g. safely hiding, on the move). By studying several past cases of terrorism and crisis response by civilians via Twitter [Beaumont, 2008; Cashmore, 2009a; Terdiman, 2009], I find that breaking news on Twitter can be attributed to the usage of mobile devices or social media publishers as civilians would be using it on the move to broadcast the situation or their current feelings/sentiments about it. This in conjunction with the findings in Dearman et al. [2008] where it is observed that users tend to share 'time-critical information' about half the time when they are in the 'mobile' state (on the move). The majority of the updates are predicted to take place when civilians are in the 'mobile' state and possibly from home Dearman et al. [2008] when civilians tend to feel safe from any direct threat.

Communication patterns	**Presence of the reply indicator (`@username message`):** which shows a strong pattern of interpersonal communication [Honeycutt and Herring, 2009] in crisis events [Hughes and Palen, 2009]. **Presence of message forwarding or retweeting (`RT @originaluser message`):** This behavior suggests the need for information sharing [Boyd et al., 2010] among civilians [Dearman et al., 2008] to disseminate more information. The dynamics of message threading and grouping behavior as above have been studied in prior work; examples are the intentions of message replying [Honeycutt and Herring, 2009], retweeting [Boyd et al., 2010], and user clustering based on communication styles [Cheong and Lee, 2010b]. Related information about the dynamics of Twitter-mediated communication is available in [Honeycutt and Herring, 2009].
Information collation	**Presence of a hashtag (`#hashtag`) signifying message grouping and categorization:** This allows easy culling of additional information by just looking up the hashtag [Cheong and Lee, 2009; Starbird et al., 2010], and allowing decision makers to know what potential aggregate information the belligerents might have.
Links to additional information	**Presence of information sharing in the message indicated by the sharing of web links (or URLs):** This complements the Twitter-based discussion with other news sources (e.g. mainstream media coverage), or simply to 'convey larger amounts of information' such as in forum posts or conventional blog posts due to the constraint of the 140 character limitation [Hughes and Palen, 2009; Starbird et al., 2010].
User-generated multimedia content	**Presence of links to user-generated content:** Links to content on sites such as *TwitPic/Flickr* pictures and YouTube videos might be a wealth of information to authorities seeking to chronicle such activities, as exhibited in prior terrorism events [Beaumont, 2008; Cashmore, 2009a] and similar disaster and crisis situations [Fleishman, 2009; Terdiman, 2009]. Information sharing behavior among users can be used to discover user-generated content about the terrorist activity (for example the extent of damage and casualties, possible identification of suspected perpetrators, newswire coverage, etc.).

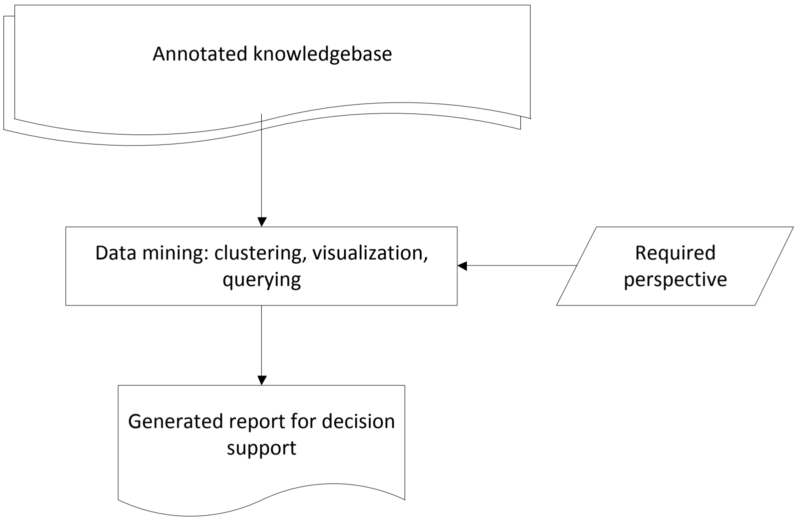

Annotated knowledgebase
Data mining: clustering, visualization, querying
Required perspective
Generated report for decision support

Figure 7.10: Framework diagram for the reporting and visualization phase (*Phase 4*).

7.2.9 *Framework Phase 4*: Pattern Detection, Visualizing, and Reporting

One of the important tasks of terrorism and disaster informatics is the mining of data to enable efficient decision-making by authorities in response to an act of terrorism. The resulting knowledgebase generated from *Phase 3* can be fed into a data mining and warehousing package.

Clustering and visualization methods can be employed to identify distinct clusters of civilians involved in the terrorism scenario based on extracted information and the integration of its desired knowledge of patterns (Figure 7.10).

For the purposes of this study, I use the self-organizing map algorithm [Kohonen, 1988] as a tool for efficient data clustering and visualization, as performed in my other published work [Cheong and Lee, 2009, 2010b; Cheong et al., 2012b]. Several other machine learning methods such as Bayesian networks, and Support Vector Machines could be employed via data mining packages such as Weka [Hall et al., 2009], but are beyond the scope of this thesis.

Also, directly from the knowledgebase itself, particular features or *perspectives* can be visualized, such as:

- **Timelines:** e.g. the rate of communication in terms of messages per unit time, time since first attack, mobility state over time.
- **Geographic heatmaps:** location of first attack ('ground zero'), distribution of tweets

These perspectives can easily be obtained simply by using data filtering tools in conjunction with spreadsheet, data analysis, and visualization packages. Examples of two such visualizations are detailed in Section 7.2.10 in detail.

The information can be further filtered in order to narrow down the scope of the required information — for example, limited to a particular slice of time, or limited to users discussing damages resulting from the incident — and also visualized in terms of charts, timelines, or social network graphs, as employed in other related research on Twitter [Huberman et al., 2008a; Java et al., 2009; Krishnamurthy et al., 2008].

7.2.10 Simulation: Synthetic Terrorism Scenario

In the real world, real terrorism scenarios are rare and unpredictable [Beutler et al., 2006]. As such, I was not able to predictively test the four-phase framework on a real-life terrorism scenario. The data for prior events such as the Jakarta bombings, and Mumbai bombings mentioned earlier could not be used, as Twitter backdates archival search (via the `search` API) to a maximum of approximately two months due to resource limitations (Section 4.1.3).

However, there is a need to test the workings of the proposed framework to demonstrate its efficacy. Therefore, I propose to experiment on the framework using synthetic datasets. The purpose of Experiment 7.6 is to simulate a terrorism-response Twitter messaging scenario involving a highly localized urban Twitter user base.

Experiment 7.6. *Preparing synthetic datasets containing metadata from Twitter for hypothetical terrorism scenarios, based on real-world events modified to include randomly distributed terrorism-related keywords.*

METHOD: As of time of writing this study in its original published form [Cheong and Lee, 2011], I selected two events for the simulation due to their nature of "mass convergence" of people [Hughes and Palen, 2009], which is likely to manifest in a hypothetical real-world terror scenario.

It is pertinent to note that none of the real-world events depicted in this experiment involve real-world terrorism scenarios; the Twitter message stream in these simulations is synthetically modified for illustrating a hypothetical scenario and it is by no means a hoax.

I create the new synthetic terrorism message data by injecting the original message content (from real-world events) with randomly distributed terrorism-related keywords as per Table 7.10 in Section 7.2.8.

The distribution of keyword injection is based on the frequency of such words existing in the 448,358-message data set [The Sunshine Press, 2009] as discussed prior (Section 7.2.8). For example, the keyword '`call`' occurs 34,552 times, with a frequency of 7%; compared to '`alert`' (5,839 times) with approximate frequency of 1%.

Again, I use the relative keyword frequency with respect to this data set due to the fact that it's a real-life capture of communication in a real terrorism scenario. All other metadata are unaltered so as not to alter the emergent user properties.

RESULTS AND DISCUSSION: The two events I have selected as the basis for simulating Twitter chatter on hypothetical terror events, and have acquired data for, are:

1. **Cuban Peace without Borders Concert/*Paz Sin Fronteras II*** (keyword: `Paz Sin Fronteras`), captured 20th September 2009. This dataset was chosen as this

real-world event is reported by Twitter users from a localized Cuban user base. Users exhibit emergent characteristics (e.g. geographic location, gender, and information sharing patterns) similar to those of real-world terrorism and crisis events which take place in a localized urban context. It is interesting to note that there are minor political controversies attributed to this concert.

2. **AFL preliminary finals**, captured 21st September 2009. This data set represents the chatter collected on the 19th September weekend where the AFL preliminary finals are being held. This dataset was chosen as it, too, had characteristics of a localized event in an urban setting. The amount of noise in this dataset is rather low; as the data is freshly harvested after the event has finished and narrowed down to the day of the event itself.

I captured 1,500 messages for both the aforementioned events via the message harvesting module of my framework in *Phase 2*. The REST-`user` API is used to query user information as detailed in Section 7.2.7 to facilitate my simple spam removal algorithm.

The details of harvested messages, including spam/noisy messages removed, are as follows (Table 7.12):

Table 7.12: Number of messages harvested for the two simulations.

Event	**Total messages**	**Removed noise**	**Sanitized messages**
Paz Sin Fronteras II	1500 (Twitter API limit)	211	1289
AFL preliminary finals	1500 (Twitter API limit)	115	1385

In essence, Experiment 7.6 is testing *Phase 2*'s data harvesting and filtering capabilities. As I am using synthetic data, I was unable to test *Phase 1*, as this simulation is operating on synthetic offline data that cannot be generated on-the-fly.

As a next part to this simulation, the demographic analyzer and sentiment analyzer in *Phase 3* is executed on my user and synthetic message dataset. The user and message metadata acquired from Experiment 7.6 is annotated via the custom-made sentiment analyzer and demographic analysis in *Phase 3*.

Figure 7.11 is a screenshot of the annotated knowledgebase resulting from *Phase 3*, imported into a spreadsheet package. From this, the properties of a collection of tweets in a quantitative form can easily be observed.

Finally, to test out and illustrate the workings of *Phase 4*, the annotated knowledgebase from *Phase 3* is then visualized and clustered to demonstrate the richness of data obtained via my proposed civilian-response informatics framework.

Two simple visualization techniques, previously alluded to in Section 7.2.9, were used to graphically illustrate data collected from the simulation in an easy to understand manner.

- **Timeline analysis:** Weka's [Hall et al., 2009] built-in analyzer is used to show the progression of the timeline with respect to user mobility state generated from *Phase*

3. This method displays states of user mobility (along the y-axis) with respect to time (the x-axis), allowing one to easily distinguish groups of people discussion the event, based on their mobility state as the event progresses.

- **Geographic heatmap:** A Google Maps 'mashup' was used to illustrate the locations of users who have contributed to chatter about a topic. This simply visualizes the location of users (who have tweets with embedded geographic coordinate metadata) on an overlay of the real topological map, making it easy to separate groups of users based on their immediate location at a particular time. Examples of similar geographic visualizations of Twitter metadata include analysis of sentiment during sports events [Bloch and Carter, 2009] and earthquake detection [Guy et al., 2010].

These visualizations are documented in Experiments 7.7 and 7.8.

Experiment 7.7. *Conducting Weka timeline analysis on simulated Twitter data to visualize user mobility.*

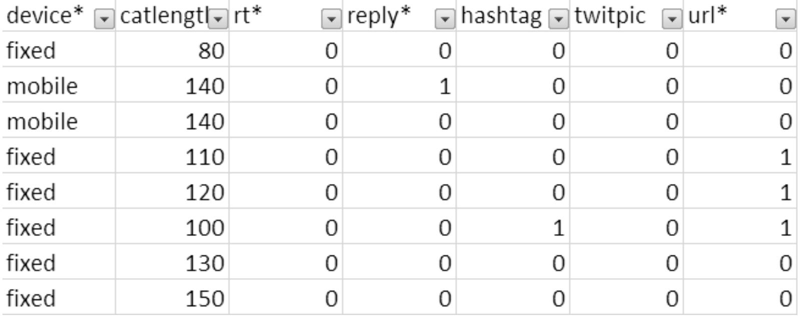
METHOD: Using the visualizer tool found in Weka [Hall et al., 2009], I was able to come up with an interesting example of timeline analysis of a terrorism event based on the findings from the obtained knowledgebase.

From the notion of *user mobility* (proposed earlier in Section 4.6.3), I use Weka's visualizer to analyze the proportion of users contributing to Twitter chatter via *fixed*, non-mobile, devices (e.g. computers, game consoles), versus *mobile* devices (mobile phones, PDAs, smart phones) over the progression of a particular event.

This is achieved by viewing the data points on the Weka visualization tool, filtered by device class, i.e. the type of `source` string for a given tweet.

RESULTS AND DISCUSSION: Figure 7.12 illustrates the distribution of messages submitted using both kinds of devices versus the progression of time. Time is measured using the messages' Unique IDs (UID), earlier found to be an effective measure of time [Cheong and Lee, 2009].

From the visualization, it can be seen that during the time of the main event, users tend to contribute less sporadically from mobile devices. This frequency tends to trail off after



device*	catlengtl	rt*	reply*	hashtag	twitpic	url*
fixed	80	0	0	0	0	0
mobile	140	0	1	0	0	0
mobile	140	0	0	0	0	0
fixed	110	0	0	0	0	1
fixed	120	0	0	0	0	1
fixed	100	0	0	1	0	1
fixed	130	0	0	0	0	0
fixed	150	0	0	0	0	0

Figure 7.11: Screenshot of the raw simulation data in the annotated knowledgebase; in this case (from left-to-right): the source Twitter client, device type, quantified message length, Retweets, replies, hashtags, presence of pictures, presence of URLs.

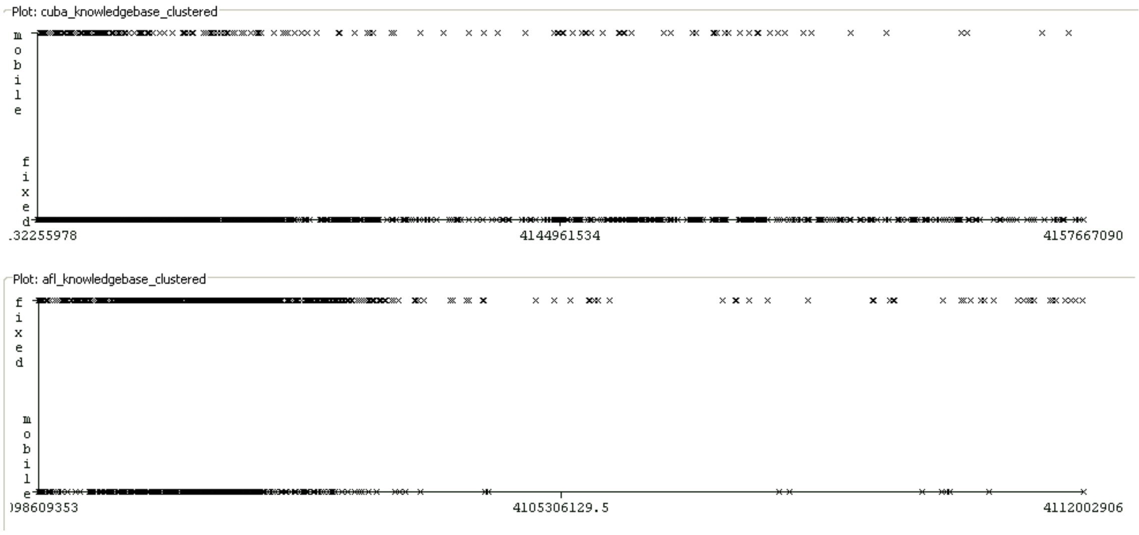


Figure 7.12: Weka timeline analysis to illustrate shift of user mobility within the *Paz Sin Fronteras II* scenario (above) and the *AFL preliminary finals* scenario (below).

the event finishes. This reflects the pattern of *information sharing* found in prior literature [Cheong and Lee, 2010c; Dearman et al., 2008; Westman and Freund, 2010; Subramanian and March, 2010; Ritter et al., 2010]. A significant number of users contribute from fixed devices, indicating that they may not be at the events themselves, suggesting that their usage of Twitter as a method of expressing views or communicating with the people on the ground.

Experiment 7.8. *Visualizing precise geographic locations of tweet authors on a Google Maps interface.*

Method: From the simulations' annotated knowledgebase, I observed that a proportion of users contributing using mobile devices have GPS-enabled mobile Twitter clients and enabled the geolocation or geotagging feature (as discussed in Section 4.6.2).

The knowledgebase is iterated to identify the records containing coordinate data. Latitude and longitude values are extracted from each record and plotted on a Google Map.

Results and Discussion: Figure 7.13 show the geolocation data from simulation scenarios plotted on Google Maps.

As can be seen in Figure 7.13, the locations of participants of Twitter conversations in the two simulation scenarios above are pinpointed using markers in the Google Maps API. This information is potentially beneficial to decision-makers and the authorities to model and chronicle civilian response to terrorist activity.

Finally, to visualize the richness of data obtained from my framework, and to extract meaningful latent patterns from them, I applied Kohonen's self-organizing map algorithm [Kohonen, 1988] on the annotated knowledgebase. A Kohonen self-organizing map [Kohonen, 1988] is a visual clustering technique projecting input from multiple-dimensions into maps of two-dimensions (cf. Section 6.1). Similar features are spatially close by on the

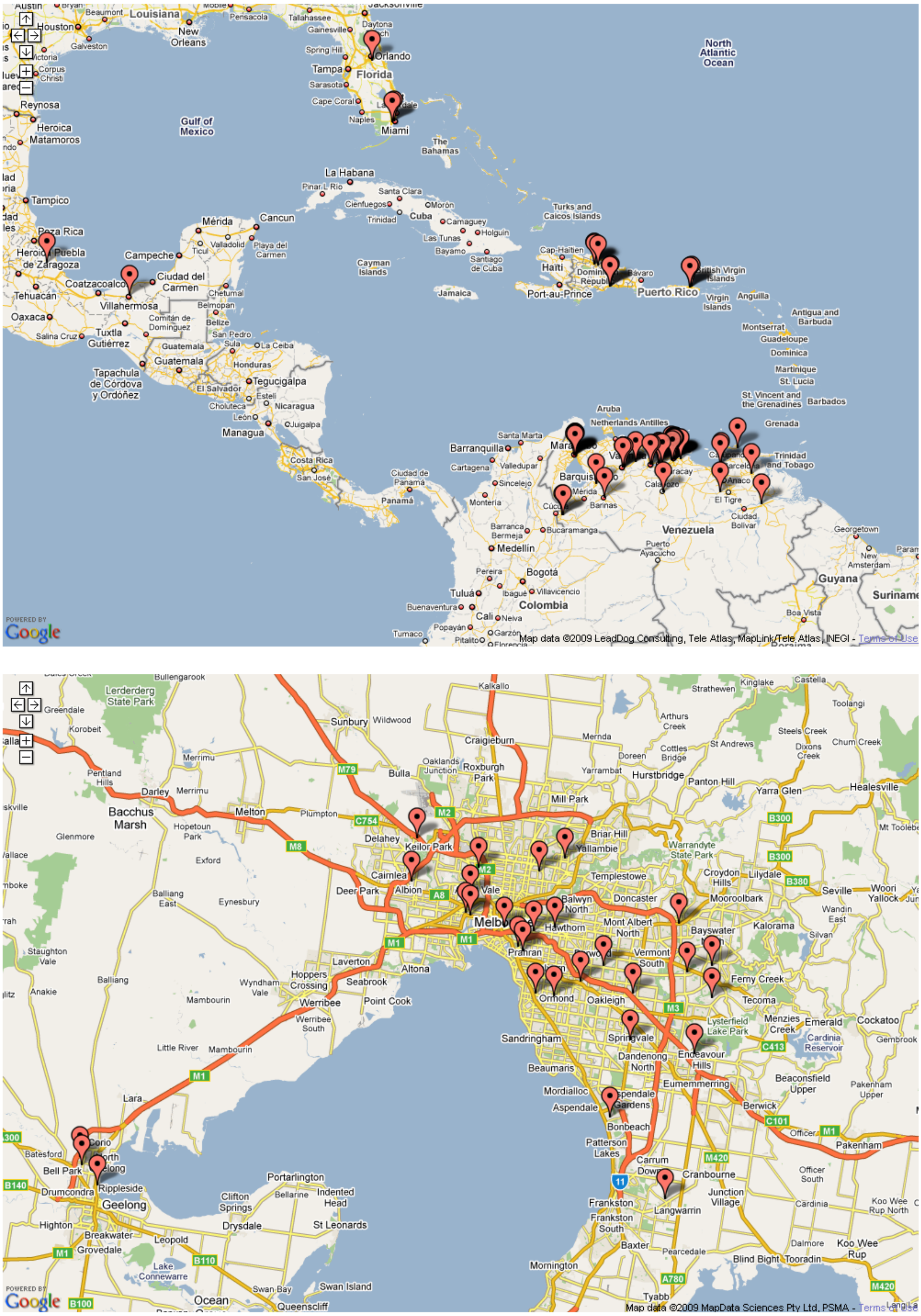


Figure 7.13: Google Maps mashup for the *Paz Sin Fronteras II* scenario (above) and the *AFL preliminary finals* scenario (below).

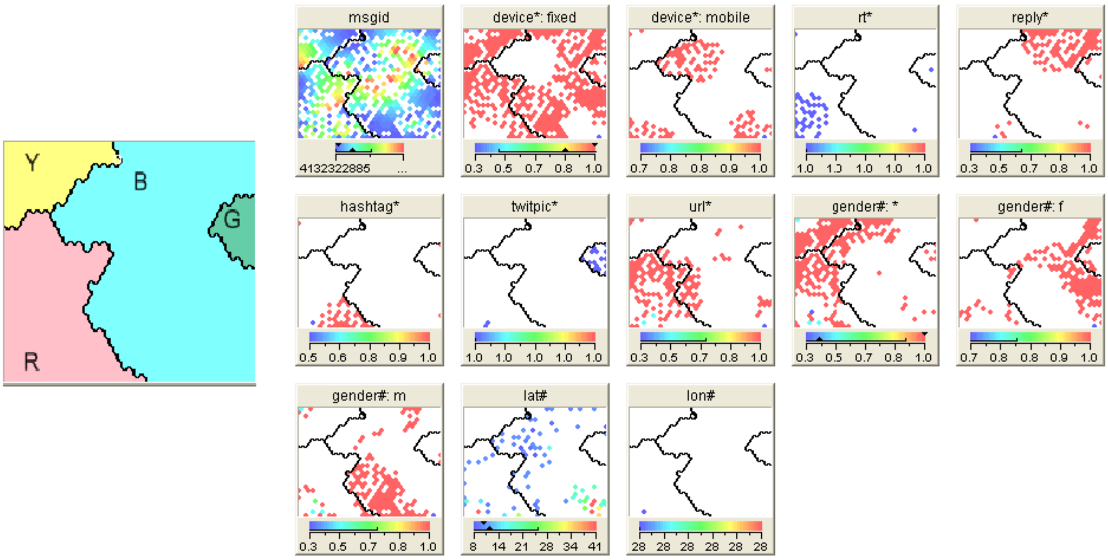


Figure 7.14: SOM clustering for the 'Paz Sin Fronteras II' simulation data.

map, which makes the self-organizing map effective for clustering and visualizing data [Kohonen, 1988; Yao et al., 2010], especially when dealing with microblog messages [Cheong and Lee, 2009, 2010b].

Clustering users based on the annotated properties from *Phase 3* can potentially reveal possible connections or similarities behind the tweets (and their authors) in an annotated knowledgebase generated by my terrorism informatics framework.

Experiment 7.9. *Performing SOM clustering on annotated records produced by the proposed terrorism informatics framework, as part of a simulated run.*

Method: I selected a subset of six demographic attributes as input to the SOM clustering algorithm: user mobility state, retweet/reply communication habits, hashtagging, photo/URL sharing, gender, and geographic location.

I chose the above subset rather than the full list of attributes, as I merely intend to study the habits and basic demography of user communication in this simulation. Bearing in mind as this is merely a simulation, I did not take into consideration the synthetically-generated data, in terms of sentiments or mood of affected Twitter users.

For each record in the annotated knowledgebase used throughout this simulation, I extracted the six attributes and fed them as input to the *Viscovery SOMine* self-organizing map package. The default attributes (as per Table 6.2 in Section 6.2) were used for this clustering exercise.

Results and Discussion: The results of SOM unsupervised clustering of the two knowledge bases generated from the simulation are as per Figures 7.14 and 7.15.

For the *Paz Sin Fronteras II* simulation, the SOM algorithm managed to segment the users into several different clusters (Figure 7.14), each with distinctive properties.

- **Blue cluster**: The majority of users, comprised of both genders, contribute from both mobile and fixed (non-mobile) devices. Their tweets contain properties of chatter (as indicated by the abundance of replies). This is visible in communication patterns involving information seeking and enquiring about the situation amongst people in the affected area; prior work discussing this 'social life' of information has been done by Hughes and Palen [2009] and Starbird et al. [2010].

- **Red cluster**: This contains hashtags and URLs in the message, reflecting the message sharing characteristics of the users, who contribute links to additional information during a potential terrorism scenario. Again, actions like these in the context of a serious event illustrate the need for microbloggers to share and receive additional information [Starbird et al., 2010].

- **Yellow cluster**: Mainly fixed device users; the genders of these users cannot be readily predictable. This might be due to the publishing of tweets by groups, corporations or agencies, which use the name of the organization as their Twitter username. This clearly distinguishes this cluster from the blue and red clusters above, as the users in the preceding clusters have readily-identifiable human names from both genders [Cheong and Lee, 2010b]. In real-world terrorism events, groups or organizations which would have a direct need for a social information presence [Cheong and Lee, 2009] would be aid agencies and organizations affected (e.g. the Hilton Hotel, which was directly sending Twitter messages offering advice to its cus- tomers during the Jakarta attack Cashmore [2009a]).

- Green cluster: users directly contributing user-generated content in the form of pictures or videos via YouTube, TwitPic or Flickr.

As for the *AFL preliminary final* simulation, the results of SOM segmentation and clustering (Figure 7.15) are interpreted as follows:

- **Blue cluster**: users from both mobile and fixed (non-mobile) devices, with tendencies to reply messages, from both genders.

- **Red cluster**: predominantly users of fixed (non-mobile) devices, having the tendency to retweet other people's messages, and share information based on #hashtags.

- **Yellow cluster**: predominantly users of fixed (non-mobile) devices, with tendency to post URLs as a method to share information.

- **Green cluster**: predominantly users of fixed (non-mobile) devices, who share information via user-generated content on sites such as YouTube, Flickr, and TwitPic.

7.2.11 Discussions, Conclusion and Further Work

In this section, I have proposed a novel framework utilizing the Twitter microblogging service as a multifaceted data source for harnessing sentimental data and demographic analysis in civilian response to terror scenarios. The novelty in this is that Twitter is rich

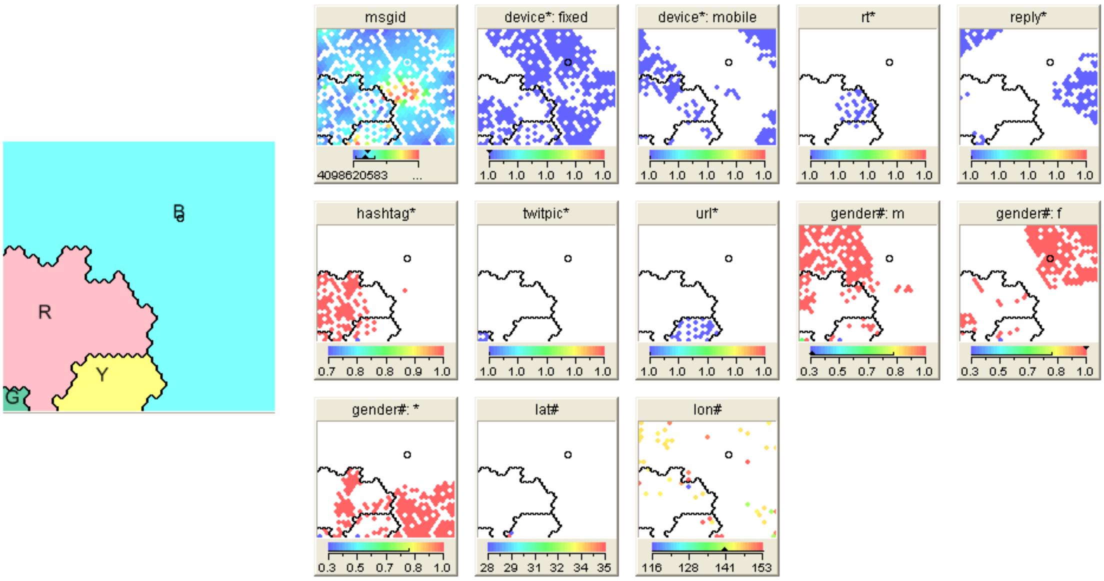


Figure 7.15: SOM clustering for the 'AFL Preliminary Final' simulation data.

in data for such applications, but not much work has been done in exposing the latent patterns and emergent properties in the context of terror informatics.

Experimental results shown here provide insight as to how my framework can be used in real-world settings by homeland security authorities and law enforcement agencies to immediately chronicle and respond to terror threats. I also shed light into the understanding of the obtained data, by coupling the harvested information with visualization and intelligent data mining techniques (SOM in my simulation example).

Due to the constraints imposed by the underlying technology this framework is built on, there are indeed several limitations that need to be discussed.

The main problem, as emphasized a few times throughout this thesis, is that the Twitter API has API limitations intended to conserve server resources [Cheong and Lee, 2010c], viz.:

- The on-demand (`search` and REST-`user`) APIs, which allow searches of historical tweets and user metadata, is heavily constrained (Section 4.1.3).

- The Streaming API, which is the API of choice for this thesis (used in the *10-Gigabyte Dataset*, Section 5.3) allows for a more thorough and a more voluminous collection of tweets. This is due to the presence of both user and message metadata in search results, and the high number of metadata records that can be fetched per unit time (Section 4.2.1). However, the Streaming API is insufficient for a near-complete retrieval of such tweets: it returns only about a 1% sample of all tweets. The ideal scenario is for Twitter Inc. to provide unlimited access to the above data, e.g. through its *Firehose* access level [Twitter Inc., 2012a]. However, this is highly unlikely due to constraints such as resources and legal issues.

There are several workarounds that can be used to overcome such hurdles. For the problem of searching for historical tweets, the workaround is to use a third-party archive of Twitter data, either by using a commercial Twitter archival service such as *TweetScan*, purchasing archived Twitter data through a reseller such as *Gnip* (for a substantially high fee), or using a corpus of ongoing Twitter captures such as the proposed US Library of Congress Twitter Archive project [Raymond, 2010]. Unfortunately, for the Streaming API, there is no other way to increase the proportion of streamed messages other than obtaining elevated privileges to the *Firehose* API access level by Twitter. As for the third problem of harvesting user information from tweets, one could utilize more client computers or use cloud computing services to perform distributed user lookups. The latter suggestion was proven feasible by Kwak et al. [2010] shortly after I originally published this case study [Cheong and Lee, 2011].

Another identified potential limitation in this framework is that some terrorism scenarios may otherwise escape automated detection. For example, if a terror threat is written in a language other than English, this framework does not fully work due to the sentiment analyzer being seeded with English key phrases. Therefore, it is preferable to include input from human observers as part of this framework to better pinpoint such false negatives. Also, it is better to have a human user tweak the framework (e.g. prioritize certain attributes in the detection such as geographic location while putting less emphasis on retweet frequency) based on the needs of the different possible terrorism scenarios.

Lastly, this study is limited due the lack of real-world data for robust analysis. *Phases 1* and *2* of the proposed framework are hard to test in a real-world situation due to the unpredictability of real-world events. Future work with this regard includes improving the validity of the aforementioned phases in the framework by deploying it on dedicated computing hardware for long-term, continuous, monitoring of the live Twitter stream.

7.3 2011 London Riots on the Twitterverse

From the previous section, I have shown how Twitter user and message metadata can easily be harvested as a rich source of data to study during times of crisis. In the final section of this chapter, I document my analysis on the 2011 London Riots as seen through the eyes of Twitter users; first published as [Cheong et al., 2012b]. This is a real-world scenario where I applied my contributions and techniques introduced in this thesis on real-world data in an attempt to 'make sense' of the London Riots.

7.3.1 Background

The 2011 London Riots are a good example of how social media and communications technology played a part in influencing, reporting, and catalyzing real-world events.

An in-depth analysis of the history, causes, and effects of the riots will be too much to detail here, but could form the bases for hypotheses to test via Twitter. To provide sufficient context, it suffices me to include a brief introduction to the riots.

The riots took place from 6–11 August 2011 (inclusive), erupting from "a peaceful protest over the police killing of a Tottenham man, Mark Duggan" [May, 2011]. On the 7th of August, the rioting, looting, and arson started in parts of London, which subsequently spread to Birmingham, Manchester and Liverpool [May, 2011]. The riots continued spreading till 10th August, where they started to ebb due in part to bad weather [May, 2011] and police response [May, 2011; Meikle and Jones, 2011]. The riots resulted in over 1,000 arrests [Meikle and Jones, 2011], a rough estimate of 100 million British pounds in damages [Lock, 2011], and a few collateral deaths [May, 2011]. Theories on the cause of the riots range from "police prejudices, a lack of social mobility, unemployment..." [May, 2011], all the way to "welfare dependency... teenage pregnancies... [and] consumerism" [May, 2011].

What is notable in the case of the London Riots is the prominent use of social media technology. BlackBerry Messenger, Facebook, and Twitter were among the oft-mentioned technologies used in the riots for "inciting public disorder" [Meikle and Jones, 2011]. "Social media such as Twitter and Facebook" were reportedly denounced as they made the "disorder in London and other UK cities" worse, according to an independent panel set up by the UK government [Halliday, 2012]. A concrete example by Adams [2011] illustrated the use of "Twitter to encourage violence" using tweets such as:

> *"Everyone up and roll to Tottenham f*** the 50 [police]. I hope 1 dead tonight"* (as quoted in Adams [2011]).

The use of such technologies, Twitter in particular, is a significant shift from traditional forms of communication during riot events, as physical presence and "shouting through a megaphone" [Adams, 2011] were necessary to organize and participate in a riot in prior years. Before the 2011 London Riots, several other riots were catalyzed by the usage of modern communication and online social media, namely the 2005 Paris riots which were partly sparked by usage of blogs [Tønnevold, 2009], and the 2005 Cronulla riots which were fueled by inflammatory text messages [Shaw, 2009].

7.3.2 Current Literature

My contributions in Chapter 4 [Cheong and Lee, 2010c; Cheong et al., 2012b] illustrated how latent metadata on Twitter can be used to provide useful inferences about the demography, online presence, and messaging habits of Twitters user base. In the previous section (Section 7.2), I have shown how Twitter can power a framework to chronicle civilian response to terrorism events, generating a wealth of information (such as mobility status of a person, and sentiments during terrorism events) from hidden metadata [Cheong and Lee, 2011].

Extant literature covered in Section 3.2.2 focused on "mass convergence [and] emergency events" [Hughes and Palen, 2009] such as disasters and mass political conventions illustrated that Twitter exhibits traits of "information dissemination... broadcasting and brokerage" [Hughes and Palen, 2009]; messages during such events exhibit information sharing (URLs) and interpersonal communication (`@user` messages). Similar research was conducted on the 2009 Canadian Red River Valley floods [Starbird et al., 2010], and tracking the geographic spread of earthquakes and forest fires [Longueville et al., 2009] over time.

Closely related to this study of the 2009 London Riots is work by Tonkin and Tourte [2012], who analyzed tweets composed during the riots (mainly in the message domain). The aims of their paper were, in a nutshell: to see if Twitter was "used as an organizational tool during the riots"; to discover motivations behind retweets; and postulate potential uses of "real-time data from Twitter" [Tonkin and Tourte, 2012]. Tonkin and Tourte [2012] posited that there was insufficient evidence to back the theory of Twitter "as a central organizational tool to promote illegal group action". Their experimental analysis showed that "irrelevant tweets died out" and that "Twitter users retweeted to show support for their beliefs in others commentaries" about the London riots [Tonkin and Tourte, 2012]. Twitter was useful as a medium in "spreading word about subsequent events" [Tonkin and Tourte, 2012], with the prevalent use of `#hashtags` to group together messages of a similar subtheme pertaining to the riots (e.g. `#OperationCupOfTea` to promote self-imposed curfews, and `#riotcleanup` to discuss about post-riot cleanup efforts).

7.3.3 Study Goals

The broad aim of my study is to analyze the commentary and participation in the 2009 London Riots as observed via Twitter, and to link observations from both the user and message domains on Twitter to real-world happenings. Specifically, the goals of this study are:

1. **Study Goal 1**: Constructing a corpus of Twitter message metadata pertaining to the London Riots, together with associated user metadata about their authors.
2. **Study Goal 2**: Inferring the demographic properties of gender and geographical location of Twitter users tweeting about the London Riots, by running my inference algorithms on the metadata.
3. **Study Goal 3**: Ascertaining presence of spatial correlations between real-life riot activity and localized Twitter chatter in England.
4. **Study Goal 4**: Characterizing user messaging intent and Twitter online presence during and after the riots. This complements the work by Tonkin and Tourte [2012] by comparing the results obtained from my dataset with their experimental results in the message domain.
5. **Study Goal 5**: Finally, discover hidden patterns by applying clustering algorithms; given the feature space of user demography, online presence, and messaging intent.

7.3.4 *Study Goal 1*: Data Collection and Sample

The Twitter Streaming API is used to collect tweets on the London Riots and their user/message metadata. The *HarvestFilter* script (discussed prior in 5.2) was used to query the Streaming API for mentions of the London Riots, and store their metadata locally for processing.

The properties of the captured data are listed in Table 7.13. For ease of reference, this dataset will be named the *London Riots Dataset* throughout the rest of this chapter.

Table 7.13: Summary of the *London Riots Dataset*, containing Twitter messages captured in the middle of the 2011 London Riots (August 9–15 inclusive).

Property	**Statistics**
Keyword filter	the hashtag `#londonriots` (similar to cf. [Tonkin and Tourte, 2012])
Data collection period	Start: Tue Aug 09 2011, 05:03:32 (UTC) End: Mon Aug 15 2011, 02:46:40 (UTC)
Number of records	503,865 messages 254,690 unique users (∼307MB uncompressed data)
Distribution of records	Aug 09 2011: 262093 records Aug 10 2011: 113188 records Aug 11 2011: 64589 records Aug 12 2011: 38997 records Aug 13 2011: 15550 records Aug 14 2011: 9061 records Aug 15 2011: 387 records

As the data capture began chronologically in the middle of the riots, I was only able to capture Twitter chatter at the tail end of the riots, including its immediate aftermath.

7.3.5 *Study Goal 2*: Metadata-based Inferences on Demography

Gender

The presence of real names on Twitter user profiles allow one to potentially infer a users gender, which is an interesting demographic property to study in relation to the London Riots, with respect to press coverage of the social dynamics of the riots [May, 2011]. Experiment 7.10 documents the process of gender classification of users in the London Riots.

Experiment 7.10. *Analyzing the Gender Distribution of the London Riots Dataset.*

Method: The gender inference algorithm, as defined in Section 4.6.1 (and tested on a large real-world dataset in Section 5.4), is applied to the *London Riots Dataset*, obtained from Experiment 7.10.

Results and Discussion: Table 7.14 details the results of the *GenderFromName* algorithm when applied to the first names found in user metadata of the *London Riots Dataset*. Out of 254,690 unique user records, there were 1,654 records with blank first names which were omitted from the analysis.

The proportion of males found in the *London Riots Dataset* exceed those of females by approximately 10 percent. This agrees with observations in news media e.g. [May, 2011] that more males participate in the riots as compared to females.

Table 7.14: Distribution of genders found from first names in the *London Riots Dataset*.

Gender	**Count**	**Percentage**
Male ♂	112,052	44.28%
Female ♀	80,417	31.78%
Unassigned	60,567	23.94%
Total	**253,036**	**100%**

Geographic location

The motivation behind studying geographic location is taken from existing studies on crisis/convergence events [Longueville et al., 2009; Starbird et al., 2010; Guy et al., 2010; Cheong and Lee, 2011] where geographic information found in Twitter user metadata forms the basis of studying how events spread spatially in the real-world.

For the *London Riots Dataset*, I apply the techniques first proposed in Section 4.6.2 and tested on real-world data in Section 5.4. This experiment and its findings are detailed in Experiment 7.11.

Experiment 7.11. *Heat-mapping to determine geospatial distribution of tweets discussing the London Riots.*

METHOD: I apply my two-phase geo-location approach (proposed earlier in 4.6.2) on the user and message metadata to determine country information for the authors of 82,049 messages.

From the 503,865 messages in the *London Riots Dataset*, I found some form of location information embedded within each message's metadata (consisting of 98,877 accurate geographic coordinates in either user/message metadata, and 404,988 free-form location strings in user metadata).

With this data, I constructed a heatmap (similar to Figure 5.4 in Section 5.4) with *OpenHeatMap* [Warden, 2012] to visualize the geographic distribution of tweets pertaining to the London Riots

RESULTS AND DISCUSSION: The results of visualizing the locations found using this experiment are per Figure 7.16.

As seen in Figure 7.16, the majority of messages in the *London Riots Dataset* originate from the United Kingdom — 47,617 tweets (a majority of 58.03 %) — as one would expect. Second to that, a proportion of messages originate from developing countries or countries with close relations to the United Kingdom, such as the United States (9.17%), Australia (3.10%), and the republic of Ireland (2.99%). The latter category of messages can potentially be attributed to Twitter users who are concerned about how the riots might affect them (or their potential spread) due to their countries' links to the United Kingdom.

From the observations on gender and geographic distribution conducted in this section, *Study Goal 2* was accomplished with the transformation of basic Twitter metadata into real-world properties of gender and geographic distribution, for characterization of Twitter observers and participants of the London Riots.

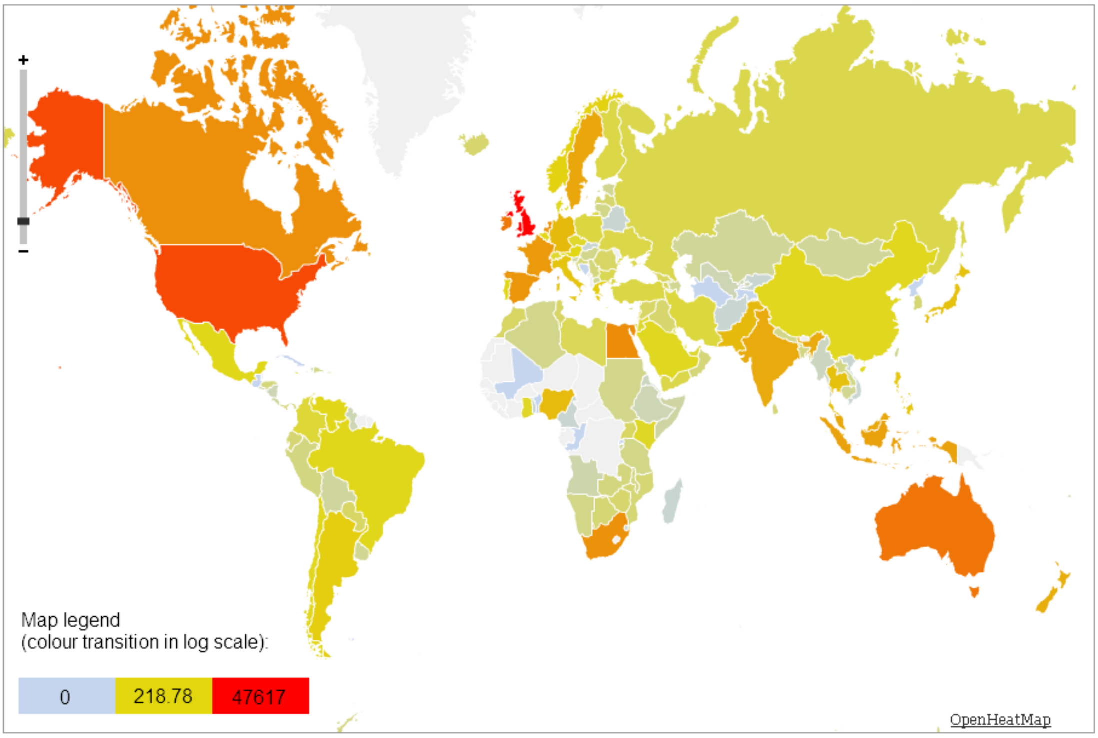


Figure 7.16: Geographic heat-map of the authors contributing to tweets in the *London Riots Dataset*. The color intensities represent the number of messages per country (scaled logarithmically). The map legend indicates the range of values represented by the heatmap colors. Generated with *OpenHeatMap* [Warden, 2012].

7.3.6 *Study Goal 3*: Correlation between Riot Activity with Tweet Locations

To visualize the Twitter activity close to the heart of the riots, I turn to ideas by James Cridland's original "Mapping the Riots" mash-up [Cridland, 2011], in which the author plotted a verifiable list of locations affected by the riots (6th August — 9th August inclusive) on Google Maps in order to visualize the riots' spatial distribution.

Cridland's original work was continued and expanded by Rogers et al. [2011] from The Guardian newspaper. After the rioting had ceased, they created a similar Google Maps mash-up using a complete list of every verified rioting incident in England [Rogers et al., 2011], as shown in Figure 7.17.

Given the availability of accurate location information (latitude/longitude pairs) in my dataset, I conducted Experiment 7.12 to accurately pinpoint clusters of Twitter activity related to the riots using a Google Maps mashup (first proposed in Section 7.2.9).

Experiment 7.12. *Determining geo-spatial distribution of tweets in the areas affected by the London Riots.*

Method: Metadata in the London Riots Twitter dataset are parsed to find accurate geographic coordinates. Of all the coordinates found, I restrict the sample space to include only coordinates which are in the United Kingdom. The coordinate bounding box for the

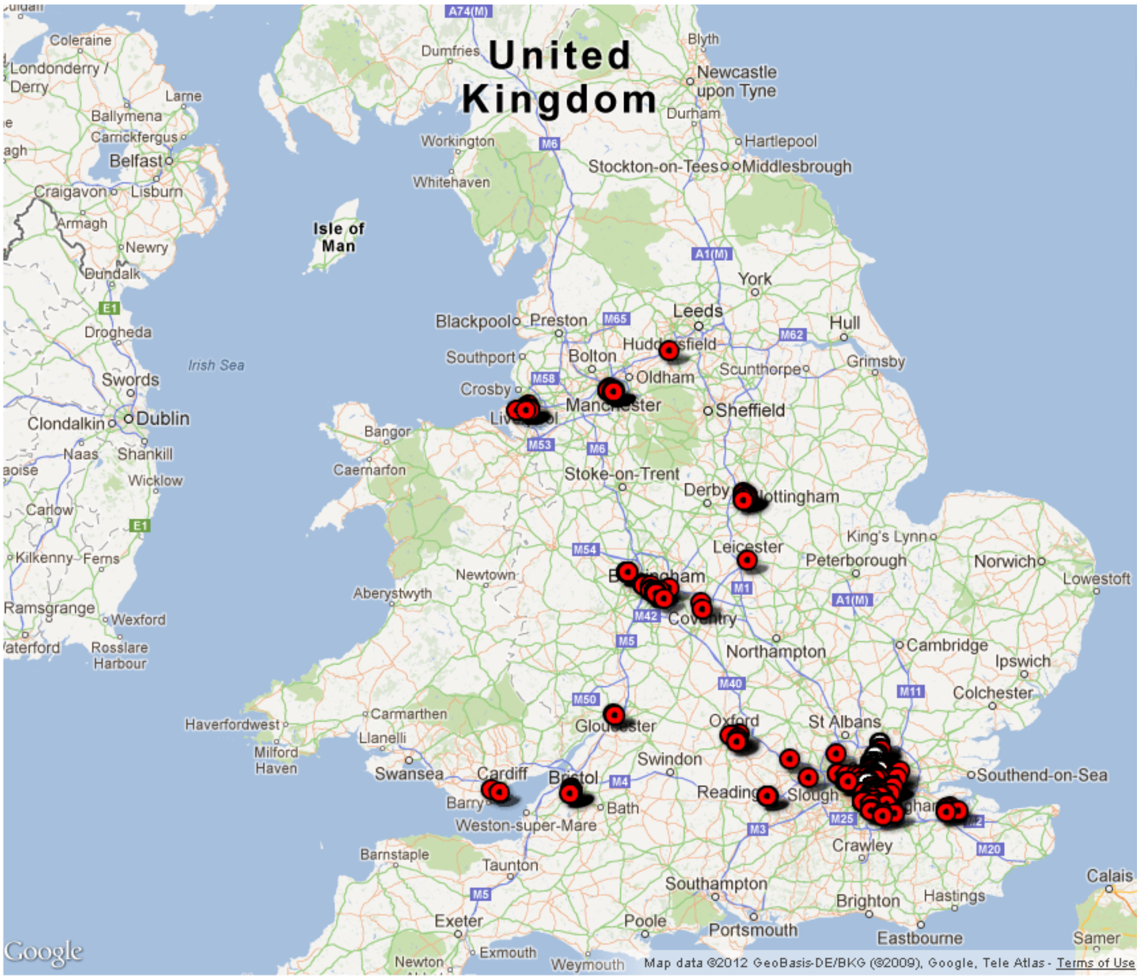


Figure 7.17: Comprehensive map from The Guardian chronicling "what has happened where as rioting spreads across England" [Rogers et al., 2011]. Each red dot indicates a reported case of rioting activity.

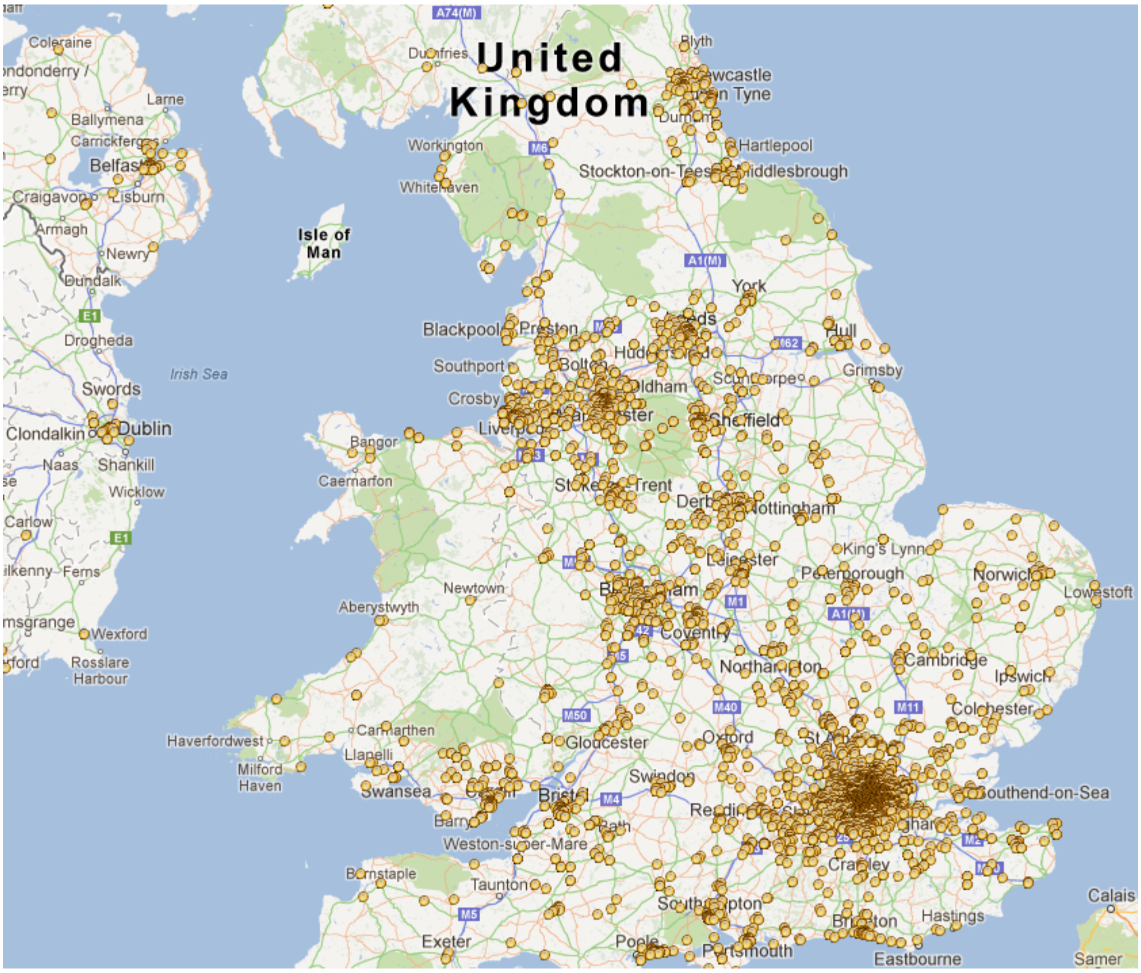


Figure 7.18: Visualization of locations found in Twitter metadata from the *London Riots Dataset*, originating from the United Kingdom. Each yellow dot represents a tweet composed in a particular location. From visual inspection, the spatial distribution of yellow dots roughly correspond to the red dots (representing riot activity) in Figure 7.17.

United Kingdom, obtained as a side effect of Algorithm 4.2, has the longitude range of (-8.1647–1.7245 degrees) and latitude range of (49.9553–60.6311 degrees).

Using a sample of 6,720 coordinates of places in the United Kingdom, I have created a similar mashup to Rogers et al. [2011] by using the found coordinates as XML input to the Google Maps API.

Results and Discussion: This Twitter-Google Maps mash-up is illustrated in Figure 7.18, where each yellow dot in the map represents one tweet.

From Figures 7.17 and 7.18, the following can be observed:

- Riot events — and correspondingly tweets — are concentrated around the most-affected areas: London, Birmingham, Bristol, Cardiff, Liverpool, Manchester, and Leeds.

- Although confirmed reports of riots are absent in major cities such as Newcastle-upon-Tyne, Southampton, and Dublin, chatter on the riots are mildly concentrated amongst these areas.

By inspection, the locations of Twitter chatter found in Figure 7.18 are close to the actual outbreaks of the riots as documented in Figure 7.17 [Rogers et al., 2011]. I conducted Experiment 7.13 to statistically test for any potential correlation between the two,

Experiment 7.13. *Testing for correlation between the frequency of tweets and frequency of documented riot outbreaks per map square.*

Method: To test for the presence of any correlation between the frequency of tweets observed per unit square on the map and the frequency of documented riot outbreaks [Rogers et al., 2011], I use the Pearson product-moment correlation coefficient [Pearson, 1895].

The map of the United Kingdom is first subdivided into squares of one degree latitude by one degree longitude. Map squares which are fully located in bodies of water, as well as those which completely fall in Ireland and Scotland, are removed. Only map squares containing parts of England and Wales are considered.

For each of the map squares under consideration, I calculate the number of tweets and the number of reported riot outbreaks which fall within. To test for correlation, I use the Pearson product-moment correlation coefficient [Pearson, 1895] (hitherto defined in Equation 5.2).

To test for statistical significance, I calculate Student's t-value [Student, 1908; Rahman, 1968], earlier defined in Equation 7.2: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ given $n =$ degrees of freedom. This test is to statistically refute the null hypothesis that there is zero correlation between the number of tweets per map square and the corresponding number of reported riot outbreaks.

As the Pearson product-moment correlation coefficient used to test for correlation (tweet count versus number of reported riot outbreaks in each map square) is sensitive to map square size, the steps above are repeated for map squares of 0.5 degree latitude by 0.5 degree longitude.

Results and Discussion: From this experiment, I have obtained the results in Table 7.15.

Table 7.15: Results of statistical tests on the total tweet count and reported riot outbreaks per unit area.

Parameter	**For 1 degree × 1 degree squares**	**For 0.5 degree × 0.5 degree squares**
Number of squares evaluated	37	114
Pearson coefficient, r	0.9704	0.9370
t-value corresponding to r (d.o.f. = number of squares-2)	23.7727	28.4027
Significant at 5% level?	Yes (exceeds required $t = 1.6896$)	Yes (exceeds required $t = 1.6586$)
Significant at 1% level?	Yes (exceeds required $t = 2.4377$)	Yes (exceeds required $t = 2.3601$)

From the results in Table 7.15, there is strong enough evidence to refute the null hypothesis that there is no correlation between tweets per unit area and the corresponding number of reported riots. It can be suggested that the correlation — between the number of tweets for a given area with the number of reported riot outbreaks in the same area — is statistically significant at the 1% level, thus accomplishing *Study Goal 3.* The results from Experiments 7.12 and 7.13 corroborates the findings from related studies mentioned earlier [Longueville et al., 2009; Starbird et al., 2010; Guy et al., 2010; Cheong and Lee, 2011] that Twitter metadata can be an accurate source of location information, useful in accurately pinpointing locations of real-world events.

7.3.7 Online Presence and Messaging Behavior of London Riots Tweet Authors

As mentioned earlier, *Study Goal 4* encompasses the study of user messaging intent and Twitter online presence during and after the riots. This is done with the tripartite analyses of (1) `source` strings and *device classes*; (2) user connectivity via the follower/friend ratio (FFR); and (3) the statistics of user-generated tweets.

Users Device Class, Mobility, and Spam

The identification of `source` strings, i.e. strings identifying the software used in composing a tweet have been covered previously in this thesis (Sections 4.6.3 and 5.4.3). Recall that *device classes* are groupings of similar `source` strings based on the platform a particular client software runs on. Experiment 7.14 documents the analysis of device classes and potential conjectures, given the `source` strings in the *London Riots Dataset*.

Experiment 7.14. *Determining device classes from* `source` *strings found in the London Riots Twitter messages.*

Method: In this experiment, I applied the methodology in Section 4.6.3 in determining device classes that are found in tweets from the *London Riots Dataset*. As evidenced in Section 5.4, this approach can potentially reveal the device classes (and in turn, user mobility and evidence of external data feeds) used in generating tweets pertaining to the riot.

Results and Discussion: Figure 7.19 illustrates the distribution of the different device classes found in the metadata of the *London Riots Dataset*. Observations from Figure 7.19 include:

- The usage of *mobile* clients outnumbered the *web* client, indicating a tendency by the participants contributing to tweets on the riots to participate while mobile or 'on the move'. This, to a certain extent, matches the observations from news reports that pinpoint mobile technology as a catalyst for participation in the riots [May, 2011; Meikle and Jones, 2011].

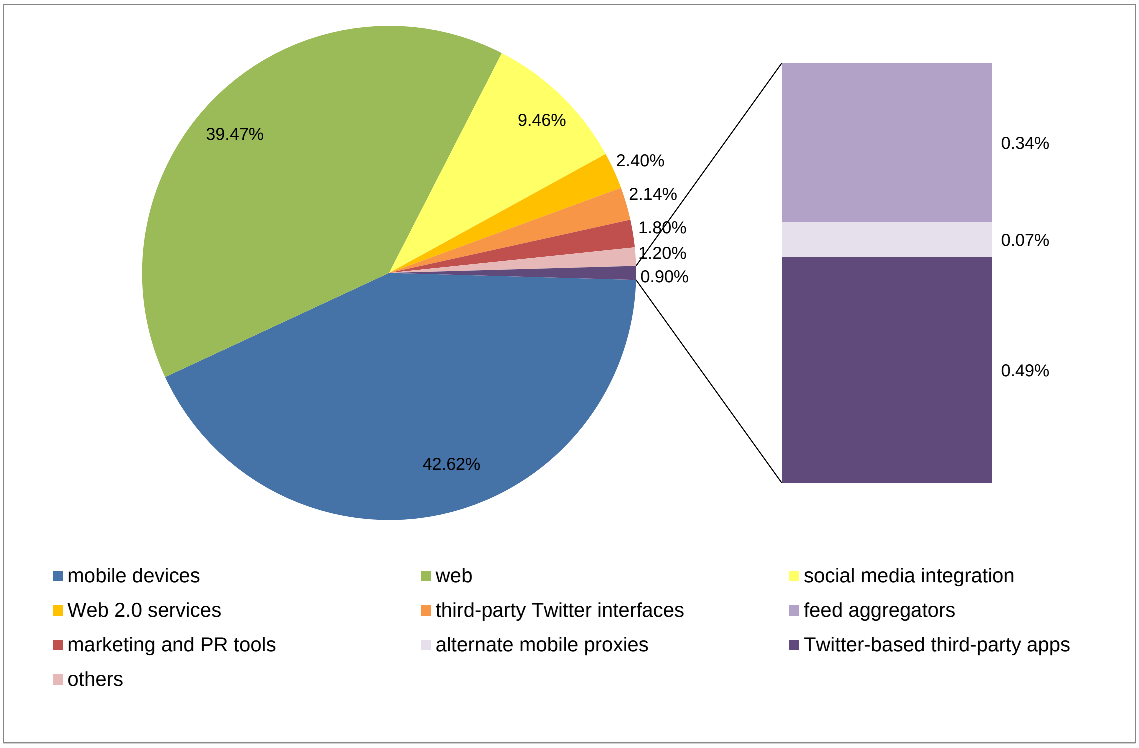


Figure 7.19: Distribution of device classes inferred from Twitter client source strings, in the *London Riots Dataset*.

- *Social media* clients and the web interface, when combined, contribute to half of the total participation during the riots. These are likely to consist of people who are not actively involved in the riots (e.g. Londoners at home or in other countries) voicing their concerns or conversing about the riots.

- The proportion of other non-prevalent device classes — such as *bots*, *suspicious/rogue applications*, *sponsored/branded Twitter clients*, and *games* — are virtually absent from the sample (and categorized under *Other* in Figure 7.18. However, this does not exclude the possibility of spammers capitalizing on the `#londonriots` hashtag to publish spam tweets.

- In the sample, there is one hitherto unseen software client, *Donate Your Account*, which is the 28th most commonly seen `source` string in the dataset (amounting to 730 tweets). This turns out to be a website which lets you 'lend' your Twitter account to a campaign, where your account will be used to broadcast tweets in support of the campaign. Upon further investigation, the accounts are 'borrowed' using this service to broadcast information by user `@CitizenRadio`, which is a political podcast.

Friend/Follower Ratio and User Influence

With the availability of user metadata in the *London Riots Dataset*, I was also able to obtain user statistics such as connectivity in terms of the Twitter social graph (Sections 4.7.3 and 5.5.3), which was absent in existing studies e.g. the London Riots messaging

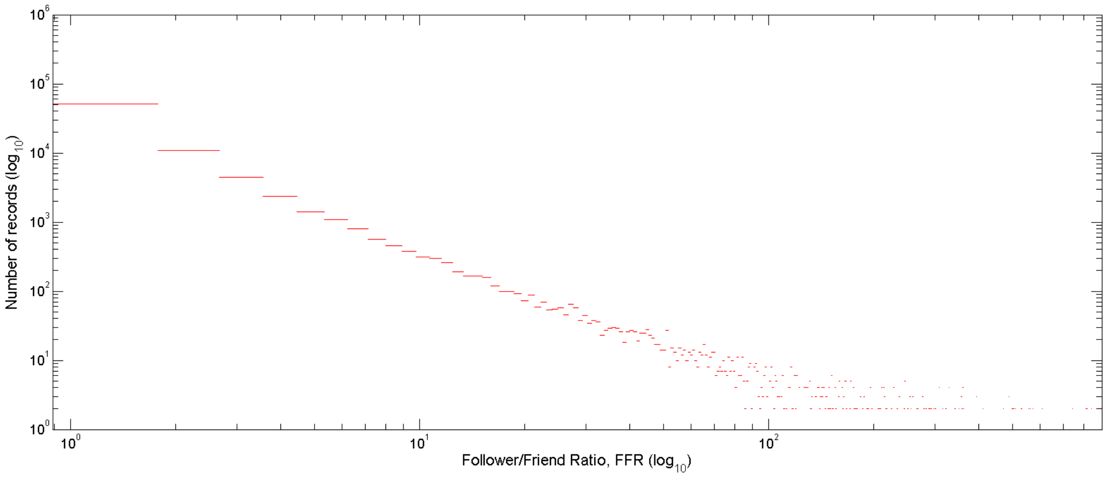


Figure 7.20: Distribution of follower/friend ratio (FFR) of unique users in the *London Riots Dataset*, sans outliers.

study by Tonkin and Tourte [2012]. Experiment 7.15 documents my study of the Twitter social graph of users in the *London Riots Dataset*, by virtue of their follower/friend ratio.

Experiment 7.15. *Analyzing the Follower/Friend Ratio (FFR) of users in the London Riots Dataset.*

Method: Using the methodology in Section 4.7.3, an analysis of the follower/friend ratio (FFR) of users in the *London Riots Dataset* is conducted. A similar analysis on the *10-Gigabyte Dataset* has been conducted in Section 5.5.3 prior.

In short, for each unique user record found in the *London Riots Dataset*, the ratio of followers to friends is calculated. A histogram plot of all FFRs in the dataset is then generated.

Results and Discussion: Figure 7.20 illustrates the FFR distribution obtained. Outliers that constitute approximately 0.04% of the dataset have been removed from Figure 7.20 for clarity.

The FFR distribution of users found in the *London Riots Dataset* follows a power law. This exhibits characteristics of a scale-free network [Barabási and Albert, 1999]. From my interpretation, users who tweet on the London Riots range from observers and participants in the riots (with a balanced number of followers), to high-profile Twitter users offering their views on the riots (with a disproportionate number of followers). This matches the FFR distribution of Twitter in general as found in the *10-Gigabyte Dataset* (Section 5.5.3).

Studying the tail end of the FFR distribution, the outliers that were removed from the graph consist of 112 records with the FFR ranging from 905.81–1,525,093.00 inclusive. Upon further scrutiny, these outliers originate from high-profile Twitter users, that can be divided into several categories:

- **Celebrities:** celebrity users, mainly from the United Kingdom, expressing their views on the riot. Examples include Jessie J, Brian Limond, and Cheryl Cole.

- **News media:** Twitter accounts by news sources and media organizations. Examples include CNN, Al-Jazeera, and the BBC.
- **Academics and writers:** high-profile academics and famous writiers offering their views. Examples include: Richard Wiseman, Andy Lorek, and Neil Gaiman.
- **Politicians:** politicians, pundits, and political sites from both sides of the political spectrum commenting on the riots. Examples include *Links Socialism*, `@ConservativeHome`, and `@anarchytweet`.
- **Satire**: notably, several members of the British Royal Family have been satirized in Twitter accounts such as `@Prince___Harry`, `@William_HRH`, and `@PrincePhilipDoE`.
- **Bots**: automated bots and algorithms that automatically retweet or post tweets about the riots. Examples include `@top_trend_world` and `@toptweets`.
- **Reactions to the riots:** several new accounts were created in response to the riots, which rapidly achieved a high FFR. Examples include `@Riotcleanup` which is a private initiative to promote cleaning up after the riots, and `@LondonRiots2011` purportedly operated by the 'British Live Information Stream' and whose intentions are unknown.

Tweet Properties and Summary Statistics

There are interesting summary statistics that can be revealed from the 503,865 tweets and the associated message metadata as per an earlier analysis in Section 4.8. In particular, in Experiment 7.16, I will be investigating the distribution of message length as an indicator of the amount of information conveyed in a tweet in the *London Riots Dataset*.

Experiment 7.16. *Analyzing the distribution of London Riots tweets' message length.*

METHOD: The length of the message string is calculated for each tweet in the *London Riots Dataset*. A simple histogram plot is then generated to visualize the frequency distribution.

RESULTS AND DISCUSSION: Figure 7.21 illustrates the generated message length histogram.

From this distribution, several observations on the messaging behavior during the London Riots can be made:

- The shortest messages found have a length of 6. These messages only contain the hashtag string `#riots`, which is the criteria for inclusion in the dataset.
- Similarly, there is a spike in the graph for length 12. These messages again contain nothing useful but a hashtag string `#londonriots`. I surmise that the only *raison d'etre* behind these tweets is for a user to 'contribute' to the overall chatter without providing any useful information.

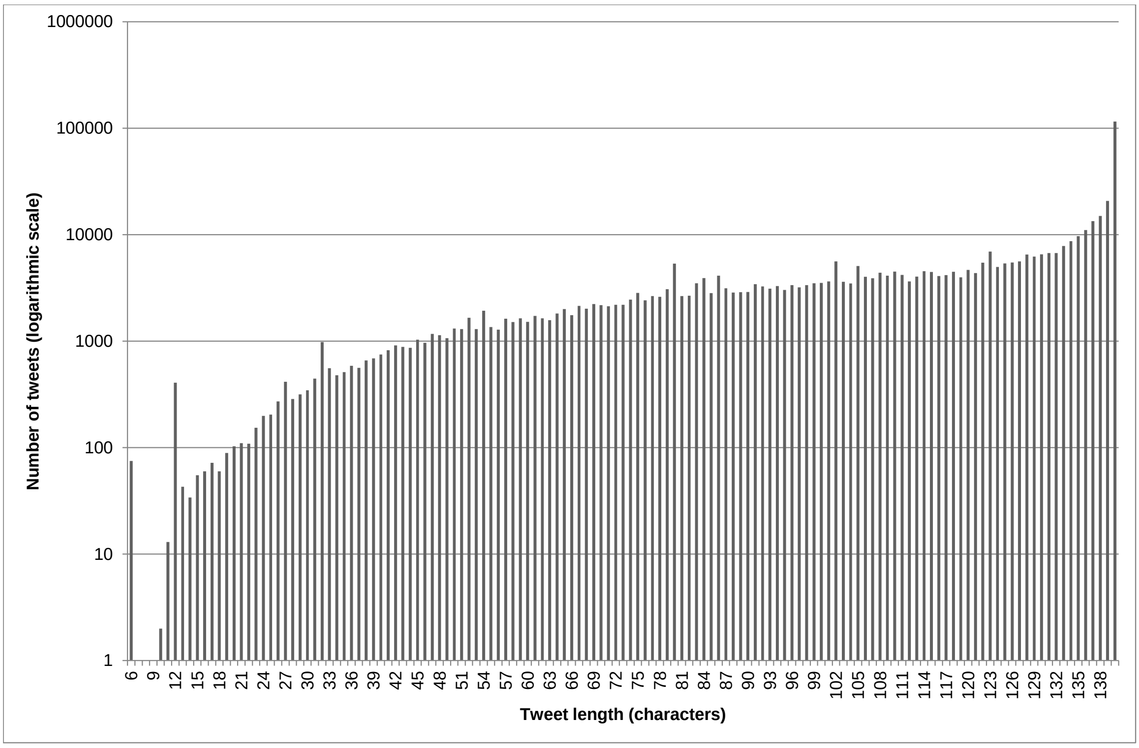


Figure 7.21: Distribution of tweet length in the *London Riots Dataset*.

- The spike of tweets at the 140-character boundary is due to truncation of long messages.

- Compared to the analysis conducted on the *10-Gigabyte Dataset* (Section 5.6) however, the histogram representation of the *London Riots Dataset* does not contain a bimodal distribution.

For the sake of completeness, I also perform a cursory examination of common entities found in London Riots tweets, to complement existing results in [Tonkin and Tourte, 2012]. The most commonly occurring Twitter accounts mentioned within tweets (in the form of @user) are from major news outlets (e.g. BBC and ITV), and campaigns to promote recovery (e.g. `@riotcleanup` as described in Section 7.3.2). As for `#hashtags`, a variety of synonyms found in the *London Riots Dataset* were used to categorize such tweets; the use of such tags corroborates the findings by Tonkin and Tourte [2012]. Place names (such as #liverpool) also occur frequently in tweets, similar to tweets found in other crisis events [Longueville et al., 2009].

To summarize this section, Experiments 7.14 characterized Twitter online presence during the London Riots via device classes, and consequently pinpointing users who are likely to be rioters. Experiment 7.15 on the other hand has shown that users in the *London Riots Dataset* follow a power-law distribution, and users with abnormally high FFRs tend to be high-profile commentators. On the other hand, Experiment 7.16 has succinctly shown that by investigating message length and frequently-occurring entities, one is able to study the messaging intent of users during the riots even without semantic/textual analysis of the contents. Together, these three experiments have accomplished *Study Goal 4*.

7.3.8 Clustering to Detect Patterns found in the Riots

After analyzing the metadata and their generated inferences from a standalone perspective in the prior section, I will now perform clustering on the variables obtained. As with e.g. Sections 6.2, 6.3, and 7.2.10, the purpose of applying clustering methods on raw (and inferred) data is to reveal any latent traits that simply aren't obvious from individual observation of the data.

For clustering, as with Experiment 7.6 (in Section 7.2.10) which tested on real-world event data found on Twitter, I use the *Viscovery SOMine* SOM clustering and data-mining package. The clustering exercise is documented in Experiment 7.17 below.

Experiment 7.17. *Performing SOM-Ward clustering on inferences generated from the London Riots Dataset*

METHOD: In the same vein as the simulations in Experiment 7.9 from Section 7.2.10 on terrorism and public response, I make use of the *Viscovery SOMine* self-organizing map package. The default attributes — as per Table 6.2, from Section 6.2 — were used for this clustering exercise, similar to previous studies in this thesis (i.e. Sections 6.2, 6.3, and 7.2.10).

The input data to *Viscovery SOMine* consisted of all 503,865 records within the *London Riots Dataset*. I have selected a set of input features for each record as per Table 7.16. All of said features were obtained using the application of inference algorithms as per Sections 7.3.5 and 7.3.7.

RESULTS AND DISCUSSION: Clustering the 503,865 records in the *London Riots Dataset* using the SOM-Ward algorithm for self-organizing map construction and visualization resulted in a total of three clusters. To make sense of the three clusters generated, the overall 2D map illustrating the final clustering is depicted in Figure 7.22.

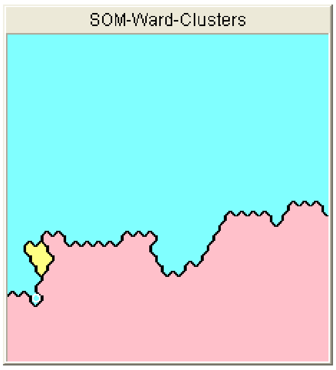


Figure 7.22: Overall clustering illustrated as a 2D map, consisting of three clusters. Cluster I (blue) is the largest cluster, occupying the top portion of the map. Cluster II (red) is the second largest, occupying the bottom of the map. Cluster III (yellow) is the smallest, and is situated on the border of the first two clusters.

Table 7.16: Features used for SOM-Ward clustering from the *London Riots Dataset*.

Feature	**Domain**	**Variable type**
Gender (as per Experiment 7.10)	User	Nominal: {*male, female, unknown*}
Country (as per Experiment 7.11)	User	Nominal: {137 unique countries, *?* (undetected), *blank*}
Device class (as per Experiment 7.14)	Message	Nominal: {13 unique devices, *other*}
Profile customization score	User	Variance
FFR (as per Experiment 7.15), log-normalized	User	Numeric
Total post count, log-normalized	User	Numeric
Message length (as per Experiment 7.16)	User	Numeric
Presence of `@user` notation	Message	Binary
Presence of `RT` notation	Message	Binary
Presence of `#hashtag` notation	Message	Binary
Presence of URL notation	Message	Binary

The three generated clusters have the following properties:

- **Cluster I**: This cluster (Figure 7.23) constitutes the majority of the records (66.09%). One feature that is of significant interest related to the cluster is the prevalence of tweets where I was not able to deduce the country of origin, possibly due to usage of invalid or fake `location` fields. In terms of device class, such messages are mainly generated using the web, mobile devices, and social media-integrated clients. The majority of records whereby no gender can be inferred from metadata (very likely due to the use of fake or non-human names) are also found in this cluster. From these features, I'm deducing that the records found in this cluster consist of 'random chatter' regarding the London Riots that originate from a wide variety of origins. The metadata provided on Twitter for these records however cannot be used to determine much about the users behind these tweets, possibly due to anonymization (e.g. invalid locations and non-human names).

- **Cluster II**: This cluster is the second largest (Figure 7.24), containing 33.75% of total records. What makes this cluster interesting to analyze is the fact that almost all tweets from Britain (country code `GB`) as well as tweets with no location information are contained within it. Narrowing down on the surface of the map where the British tweets are concentrated, there is a visibly significant proportion of male users, again agreeing with the results of Experiment 7.10 conducted earlier. In the same region (as with the British tweets), the distribution of message lengths tend toward the 140-character limit, suggesting a high information content per tweet. Device classes found in the same map region tend to be either web or mobile devices. The prevalence of mobile devices in this cluster agrees with Experiment 7.14, where

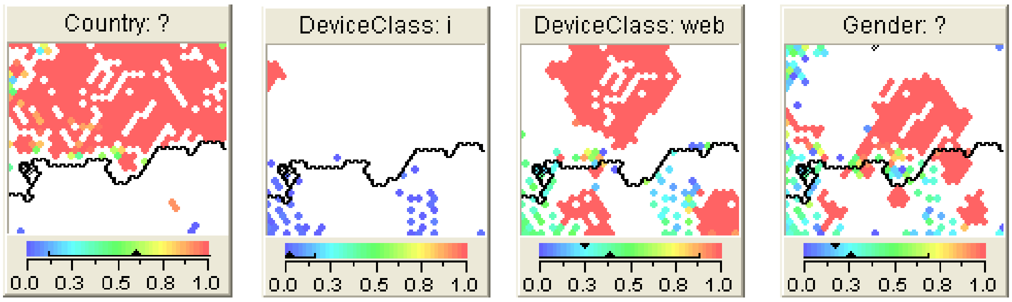


Figure 7.23: Features of interest for Cluster I. Data points residing in this cluster exhibit characteristic values (bright red-colored regions) cf. said features. Maps (from left-to-right) indicate: a large presence of invalid `location` fields; the exclusive use of social media clients by users within this cluster; a high proportion of web clients originating within this cluster (red values); and a large presence of fake/non-human names in this cluster (no gender was deducible from such names).

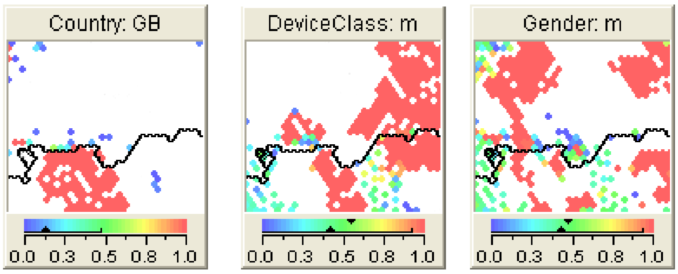


Figure 7.24: Features of interest for Cluster II. Data points residing in this cluster exhibit characteristic values (bright red-colored regions) cf. said features. Maps (from left-to-right) indicate that: this cluster contains almost all the tweets geolocated to Britain (`GB`); mobile devices form a proportion of tweets in this cluster; and this cluster contains many male users (as inferred via first name).

they are identified as catalysts for participation in the riots [May, 2011; Meikle and Jones, 2011].

- **Cluster III**: This cluster is the smallest (Figure 7.25), comprising merely 0.16% of the overall input data size. However, this cluster, which appears to be anomalous with respect to the rest of the clusters found, exhibits a quaint property in terms of the origin of tweets. Tweets contained within this cluster entirely originate from the following list of countries — the United Arab Emirates, Greece, Brazil, Qatar, Syria, Trinidad and Tobago, and Zimbabwe. From the list, Brazil, Syria, Trinidad and Tobago, and Zimbabwe are developing nations; while the United Arab Emirates and Qatar are wealthy Middle-Eastern countries. This might suggest that users participating in Twitter discussions on the London Riots from these countries share a common concern between them about these riots.

7.3.9 London Riots: Case Study Conclusion

In essence, the entire study on the 2009 London Riots in this section has shown that analysis of social metadata can yield useful insights about major social events.

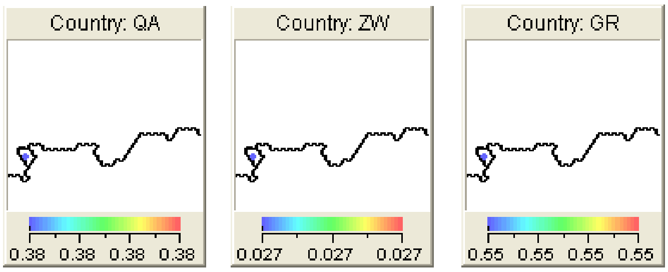


Figure 7.25: Features of interest for Cluster III. Data points containing such features are unique to Cluster III, albeit with a small proportion illustrated by a small purple-colored region within said cluster. Maps (from left-to-right) indicate: users from Qatar; users from Zimbabwe; users from Greece.

Major conclusions drawn from the study include:

- **Gender dynamics in the riots**: There are more males than females in London Riots tweet authorship; in contrast to the pattern of female-majority in general (as per Section 5.4.1).

- **Real-world rioting versus tweeting activity**: A huge statistical correlation was found between the tweet origins and the real-world riot locations.

- **Efficacy of mobile communication in riots**: The proportion of tweets from mobile devices was the highest, suggesting a possibility of their catalyzing riots.

- **Awareness on the riots on Twitter**: From analyzing FFRs and message summaries from high-profile Twitter users, another segment of tweets that are focused on commentary and recovery initiatives were detected.

- **Latent patterns in riot tweets**: By clustering the user and message properties using Kohonens SOM, three clusters were obtained; each one of them exhibits unique spatial and behavioral characteristics.

Further experimental investigations are required to draw decisive conclusions on behavioral patterns, with emphasis on their clustering.

7.4 Concluding Notes

In this chapter, I have performed three studies, two of which are based on real-world Twitter data from real-world events — the 2009–2012 Earth Hours and the 2009 London Riots — and the other, a theoretical framework for potential chronicling of terrorism events.

These studies have illustrated the efficacy of my approaches in Chapter 4 in analyzing and distilling raw Twitter metadata from both user and message domains, for the study of real-world, high-impact phenomena. This goes to further show that Twitter can effectively 'mirror' the real-world, and can complement existing approaches to studying phenomena such as riots, activism, and terrorism from the arts and sciences.

In the penultimate chapter, I will detail several eclectic approaches to the analysis of Twitter data which has been conducted throughout the course of my research. These novel ideas deal with trend analysis (with respect to Twitter Trending Topics), and a simple approach to measuring and qualifying trend persistence using the deprecated Twitter On-demand (REST-`user` and `search`) APIs.

Chapter 8

Eclectic Approaches to Twitter Data Analysis

> "*...and I try not to dream,*
> *but them possible schemes...*
> *swim around, wanna drown me in sync*"
>
> — Norah Jones
> *Chasing Pirates* (2010).

Parts of this chapter have been published as:

Cheong, M. [2009]. What are you Tweeting about?: A survey of Trending Topics within the Twitter community, *Technical Report 2009/251*, Clayton School of Information Technology, Monash University.

Cheong, M. and Lee, V. [2009]. Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base, Proc. CIKM 2009 Co-Located Work- shops: SWSM 2009, pp. 1–8.

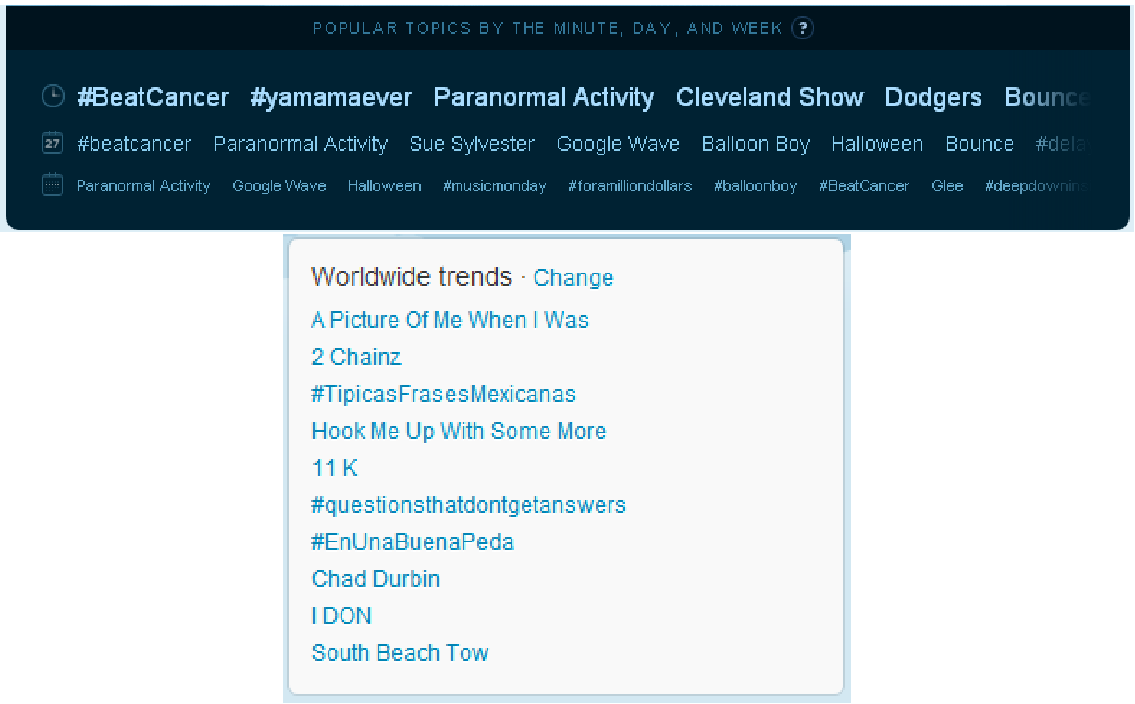
In the penultimate chapter of this thesis, I will showcase several eclectic techniques of Twitter metadata and trend analysis. The results presented in this chapter are side-effects from my overall PhD research, that are nonetheless novel and worthy of introspection.

This chapter is divided into two parts: the first being methods to examine Twitter trends via the Trending Topics feature (with its own API). This first part contained within Section 8.1 focuses on trend-based metadata collection using the on-demand APIs (`search` and REST-`user`); how crowd-sourced interpretations of memes and trends can be used to understand a given topic which is peaking in popularity on Twitter for a given timeframe; and how such trend-based analyses can complement the dual-domain metadata inference and clustering methods that form the bulk of this thesis.

The latter part of this chapter contains secondary results from analyses of Trending Topics in terms of trend persistence. I have conducted such analyses in the early part of my PhD. These results have since been superseded and independently vindicated by later work, but do serve as proofs-of-concept which supplement my findings from the rest of this thesis.

8.1 Trending Topics and User Behavior

Twitter allows users to observe the top ten popular terms or topics of discussion at any given moment through its ‘Trending Topics’ feature (Chapter 2). One of the interesting features of Twitter since its inception is the existence of Trending Topics — a list of top ten most tweeted topics — ranked by Twitter’s proprietary algorithm (Figure 8.1).



POPULAR TOPICS BY THE MINUTE, DAY, AND WEEK ?

#BeatCancer #yamamaever Paranormal Activity Cleveland Show Dodgers Bounc

27 #beatcancer Paranormal Activity Sue Sylvester Google Wave Balloon Boy Halloween Bounce #dela

Paranormal Activity Google Wave Halloween #musicmonday #foramilliondollars #balloonboy #BeatCancer Glee #deepdownin

Worldwide trends · Change

A Picture Of Me When I Was

2 Chainz

#TipicasFrasesMexicanas

Hook Me Up With Some More

11 K

#questionsthatdontgetanswers

#EnUnaBuenaPeda

Chad Durbin

I DON

South Beach Tow

Figure 8.1: Sample screenshots of Twitter Trending Topics accessible via <http://twitter.com/>. Apart from website design differences and the option to localize Trending Topics to one’s local area, the concept behind Trending Topics have not changed throughout the years, as can be seen in (a, above) screenshot circa 2009 and (b, below) screenshot circa 2012.

The Trending Topics list is a useful feature to discover the topics of interest to the Twitter community; in a way, measuring the collective and emergent behavior exhibited by Twitter users worldwide at any given moment.

I have conducted a case study on Trending Topics, originally published as [Cheong, 2009]. Within this section (Section 8.1), I will document my findings with regards to a high-level approach in studying Trending Topics.

8.1.1 Case Study Goals

My case study of Trending Topics [Cheong, 2009], covered in this section, has the following four goals:

1. Designing a methodology for harvesting of Trending Topics, sanitization of acquired data, and its archival.

2. Using crowd-sourced interpretations for annotation and explaining of Trending Topics.

3. Attempting to localize/pinpoint the geographic nature of Trending Topics, given their crowd-sourced annotation.

4. Analyzing the statistics on trending behavior in terms of significance of Trending Topics among the Twitter user base, and observing the evolution of these Trending Topics over time.

This case study draws its inspiration from the demographic analyses performed in my earlier paper [Cheong and Lee, 2009]. Compared to [Cheong and Lee, 2009], the study in [Cheong, 2009] is a high-level, 'big picture' overview of the nature of Trending Topics on Twitter.

In the non-academic domain, coverage on Twitter Trends in mainstream and popular media has been steadily increasing. To illustrate this, a simple search on Google News (conducted on December 3 2009 via <`http://news.google.com/news?q=twitter+trends`>) for the key phrase `Twitter Trends` returned 1,661 unique news entries. One such example is from *The Independent* — a UK-based newspaper — which has a weekly online feature dissecting each week's popular trends [Relax News, 2009].

8.1.2 Goal 1: Trending Topics Acquisition and Archival

Software

For the first goal, I begin by describing the development of a simple 'listening post' Java program based on the Twitter REST API [Twitter Inc., 2009]. I was limited to using the older REST API as this study was conducted in late 2009, before development of the Streaming API has matured.

In theory, the program polls the Twitter API to retrieve a list of ten Trending Topic strings after a specified time interval. I chose five minute interval to avoid wasting resources and overloading server traffic, while at the same time providing sufficient granularity in observing the trends' movements. This abides by the older REST API restriction on API calls, so as to not abuse white-listing permission granted by Twitter for my research in 2009–2010.

The 'listening post' records every trend as reported verbatim by the API, then marks its minimum and maximum position on the Trending list ($1 - 10$ inclusive). To track its permanence on the Trending list, a simple counter mechanism is implemented. After every 5-minute poll of the API, the counter is incremented; and if the trend is seen for the first time, it is timestamped.

Data Sanitization

The data is then sanitized to merge duplicates caused by irregular casing. An important point to note is that from my observations in this case study, Twitter's proprietary trending algorithm exhibits case-sensitivity with respect to the strings found. Some of the Trending Topic strings have differing case in between different API polling results — e.g. from `lowercase` to `CamelCase`.

Obtained Data

The 'listening post' was run for ten days, from 11 — 21 November 2009 inclusive. The results obtained at the end of the observation period stored in a CSV file for further analysis. Over the observation period, I have collected 677 Trending Topics[1]

The only preprocessing performed on the strings are merging the different variations in case. As a result of the sanitization process, 466 unique topic strings are obtained. This obtained set of sanitized topic strings shall be referred throughout this case study as the *Trending-Topics-Nov2009* dataset.

Note that similar strings (distinguished based on their spelling and phrasing) are *not* collated together, as the purpose of this study is to survey the obtained trends without performing additional assumptions e.g. disambiguating terms or collating them to a common subject.

Following from this, several points can be observed about the behavior of Twitter's Trending Topics algorithm. These weaknesses, evident in the Trending Topics strings observed towards the end of 2009 — when this case study was conducted and published [Cheong, 2009] — can be alleviated somewhat by using text post-processing algorithms. Though beyond the goals of this case study, I also suggest potential post-processing methods that can improve the quality of the Trending Topic strings.

1. It does not automatically group together a string and its related hashtag [prefixed with a hash (`#`) symbol].
   Example: '`#oprah`' and '`oprah`' are treated as separate Trending Topics, not as one single topic.
   Potential solution:: Performing string-matching by disregarding punctuation marks and other non-alphanumeric characters.

2. Different variations in phrasing the same subject or typographical differences create two or more Trending Topic strings.
   Example: "`chrome os`", "`google chrome os`", "`with chrome os`", "`google chrome os to`" [2] are four separate topics.
   Potential solution:: Using n-gram similarity to match common groups of words, or n-grams, in Trending Topic strings. This is a common technique used in social media mining [Russell, 2011b].

3. Due to the above, certain trend strings have no particular meaning until it is rephrased in the context of the original Twitter messages.
   Example: the keyword '`lax`' which was observed in the *Trending-Topics-Nov2009* dataset does not refer directly to the LAX Airport as a subject, nor the English

[1] The data collection was briefly interrupted for at most a few hours during the course of data collection due to network issues and Twitter scheduled maintenance; the 'listening post' program was resumed immediately after connectivity was established.

[2] The last phrase was originally '*Google Chrome OS To Launch Within A Week*', quoted verbatim from TechCrunch, an influential technology blog. The process of quoting the phrase verbatim (a 'retweet' or RT in Twitter) caused it to be a separate Trending Topic. The original URL written by Michael Arrington is at <`http://www.techcrunch.com/2009/11/13/google-chrome-os-to-launch-within-a-week/`>.

adjective 'lax' [i.e. not strict]; it actually refers to news of Mike Tyson arrested at LAX airport.

Potential solution:: Again, using n-grams to provide context by listing words that frequently co-occur with a given Trending Topic string [Russell, 2011b].

Since the publication of my case study [Cheong, 2009], however, Twitter has vastly improved their Trending Topics algorithm, and has eliminated the three weaknesses listed above. A cursory inspection of Trending Topics conducted in mid-2012 has confirmed this fact.

Prevalence of Hashtagging Behavior

Hashtag use, as shown in my Section 4.8.2 on metadata-based message inferences, is an indicator of social tagging to categorize posts. Social tagging exists to allow ease of communication and searching for related posts (further discussion on hashtags can be found in [O'Reilly and Milstein, 2009], [Cheong and Lee, 2009], and [Boyd et al., 2010]). Given the obtained *Trending-Topics-Nov2009* data, I was easily able to perform a quick study on the proportion of hashtags among Trending Topics (Experiment 8.1).

Experiment 8.1. *Analyzing the proportion hashtags among the Trending Topics in the Trending-Topics-Nov2009 dataset.*

METHOD: Simply counting the number of hashtagged Trending Topics, which are one-word strings prefixed with a hash symbol (#).

RESULTS AND DISCUSSION: Table 8.1 illustrates the proportion of Trending Topics containing hashtags, which amounts to about 30%.

Hashtag presence	**Percentage**
Hashtag present	28.76% (134 out of 466)
No hashtag	71.24% (332 out of 466)

Table 8.1: Presence versus absence of hashtags in *Trending-Topics-Nov2009.*

In my interpretation of this, social tagging and self-organizing behavior is present among the user base contributing to the discussion of a Trending Topic to a certain extent. As discussed in [O'Reilly and Milstein, 2009], hashtags are a user-created "...ad hoc solution. . . to categorize a message", as historically, there was no tagging system in Twitter before the introduction of hashtags.

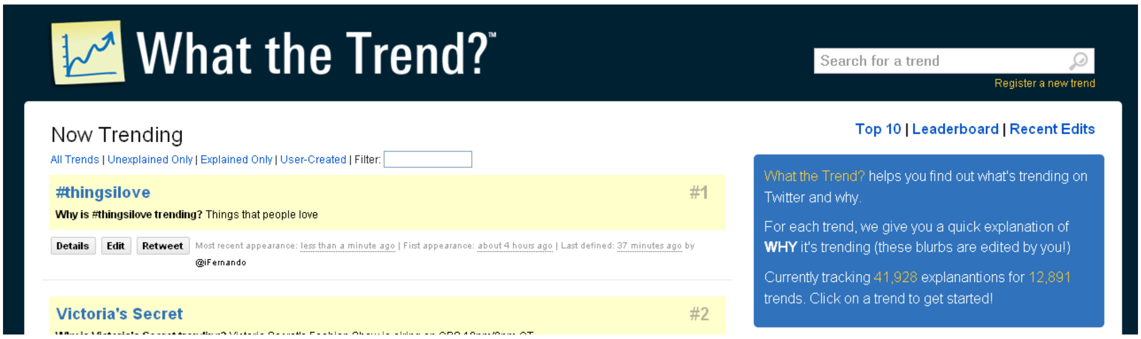
Once the list of trend strings have been sanitized and annotated, they are then imported into a spreadsheet for further analysis (i.e. Goals 2, 3, and 4 of this case study).

As this is a high-level exploratory survey on the trends themselves, no user demographic information will be obtained from the metadata of the user base and their messages, in contrast with research such as [Cheong and Lee, 2009], [Java et al., 2009], and [Huberman et al., 2008a].

8.1.3 Goal 2: Crowd-Sourced Interpretation and Categorization

The second goal behind this study is the annotation of the acquired Trending Topic data using crowd-sourcing....

The noun, concept, or meme behind each trend string is then interpreted using *What The Trend?* [Mayer, 2009], a collaborative website designed for users to describe and explain the meaning behind each trend string (Figure 8.2).


What the Trend?™
Search for a trend
Register a new trend
Top 10 | Leaderboard | Recent Edits
All Trends | Unexplained Only | Explained Only | User-Created | Filter:
Details
Edit
Retweet

Now Trending

#thingsilove #1

Why is #thingsilove trending? Things that people love

Most recent appearance: less than a minute ago | First appearance: about 4 hours ago | Last defined: 37 minutes ago by @iFernando

Victoria's Secret #2

What the Trend? helps you find out what's trending on Twitter and why.

For each trend, we give you a quick explanation of **WHY** it's trending (these blurbs are edited by you!)

Currently tracking 41,928 explanantions for 12,891 trends. Click on a trend to get started!

Figure 8.2: *What The Trend?*, a crowdsourced trend description page, available at <`http://www.whatthetrend.com/`>.

What The Trend? is used to provide crowdsourced interpretations of the trend strings, as the official Twitter site itself [Twitter Inc., 2009] utilizes *What The Trend?* to explain the trends on their website (as of 2009).

For trends that have no explanation, Google (specifically its Translate and News services) is consulted to provide interpretations to trends in other languages or trends that are not defined on the *What The Trend?* website — e.g. by translating the string from a different language into English, and searching for proper context in newswire services.

There is no standard set of category labels that can be directly used for Twitter trend data, as categorization of trends on Twitter have not been given much emphasis as seen in extant work.

On the other hand, there exists a (non-exhaustive) listing of categories used by studies on social awareness [Naaman et al., 2010]. However, these categories are more focused on the information needs of an individual in the domain of individual computer-human interaction [Naaman et al., 2010], which is unsuitable for this categorization task.

Therefore, I have decided to create a list of categories tailored to classify the obtained trends. Information sources and prior research used to develop this list include:

1. Tags available on *What The Trend?* [Mayer, 2009]

2. The general categories of topics surveyed by Cheong and Lee [2009] in their original Twitter Trends research paper.

3. Habits of information sharing, cf. Dearman et al. [2008].

4. The habits of live reporting, cf. O'Reilly and Milstein [2009] and Ebner and Schiefner [2008].

5. Internet memes and viral information sharing, cf. Arbesman [2004], Wasik [2009], and Hodge [2000].

The final list of categories used to annotate the tags are as per Table 8.2.

From the *Trending-Topics-Nov2009* collection of Trending Topics, I will now focus my attention on categorizing them (Experiment 8.2) based on Table 8.2.

Experiment 8.2. *Categorizing the Trending Topic strings found in Trending-Topics-Nov2009.*

Method: Using the descriptions per Table 8.2, I assign each string in *Trending-Topics-Nov2009* to an appropriate category. Crowd-sourced definitions and interpretations of trend strings from *What The Trend?* are used whenever possible; failing which a Google search is used to determine the nature of a trend string.

Results and Discussion: Based on the annotated categories in 8.2, I obtain the following category distribution of Trending Topic strings (Figure 8.3).

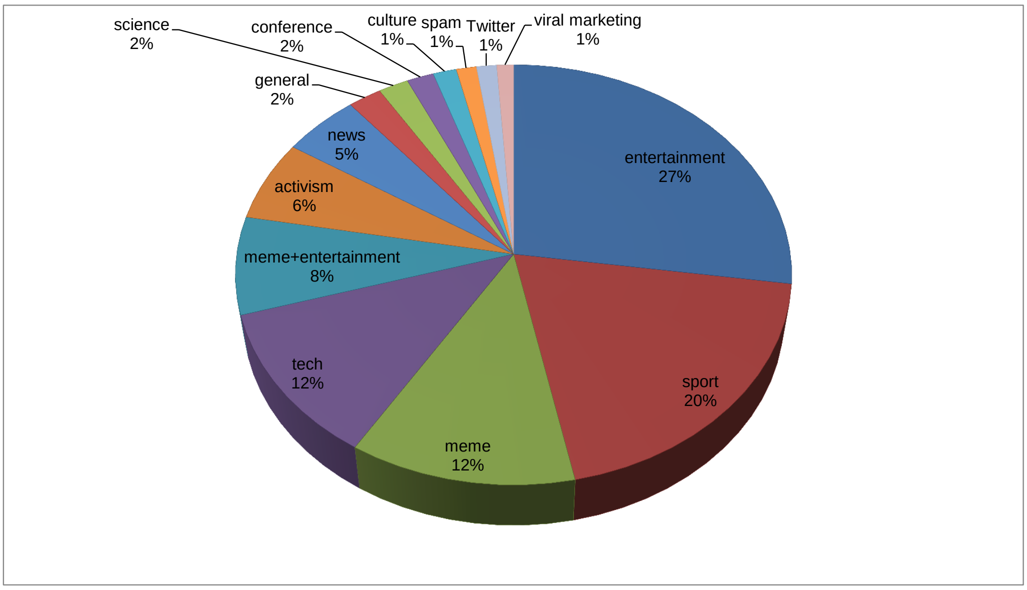


Figure 8.3: Breakdown of Trending strings by category in the *Trending-Topics-Nov2009* dataset (rounded to the nearest percentage point).

The top four trends surveyed account for over 70% of all trends; these are, in descending order:

1. entertainment-related topics (~27.25%, or 127 trends)
2. sports news (~19.53%, or 91 trends)
3. Internet memes (~12%, or 56 trends)
4. technology news (~12%, or 54 trends).

Memes that have their origins in entertainment (e.g. started by celebrities, or about a particular artist) account for 35 trends (7.51%). If this *memes in entertainment* category

Category	**Trend refers to**
activism	usage of Twitter for activism, such as to spread awareness about a charity. (Example: `#worlddiabetesday` for awarness of World Diabetes Day on 14 November 2009).
conference	users 'conference-Tweeting' about an ongoing conference. (Example: The TEDx conference in Amsterdam with its hashtag `#tedxams`).
culture	popular culture (Example: `jimmy choo` refers to a designer shoe brand).
entertainment	entertainment, e.g. music, television, movies and celebrities. (Example: `#newmoon` refers to the sequel to the Twilight movie series).
general	common phrases and proper nouns, but without sufficient context to frame the trend. (Example: `goodmorning`, which literally refers to the everyday greeting. This keyword was used in an earlier experiment, cf. Experiment 4.10).
meme	Internet and Twitter-based memes. (Example: `#musicmonday` is an old Twitter meme where users exchange music recommendations).
meme+entertainment	memes that originate from popular entertainment or were started by a celebrity. (Example: `#weloveyoujustin` was a Twitter meme created by fans of Justin Bieber the singer).
news	current affairs and local/global news, which includes crisis events. (Example: `richard heene` refers to the perpetrator of the Balloon Boy hoax in October 2009 that became a news headline).
science	scientific news, excluding IT/technology-related subjects. (Example: `#nasatweetup` refers to NASA's Twitter 'tweet-up' where it engages members of the public in conversation).
spam	spam, phishing attempts, malicious activity. (Example: `and lots more` is a phrase found in Twitter spam towards the end of 2009).
sport	sporting events and sports news. (Example: `manny pacquiao` refers to a professional boxer).
tech	specifically IT/technology-related subjects such as games, gadgets, and software. (Example: `google chrome os` refers to a cloud-based operating system by Google).
Twitter	official changes to the Twitter service introduced by Twitter Inc. (Example: Twitter maintenance occuring in the `pacific` timezone).
viral marketing	use of Twitter to virally promote a product, without malicious intent, in contrast with the *spam* category (above). (Example: Comet Group Ltd., a UK-based appliance store, promoting sales with the hashtag `#cometparcel`).

Table 8.2: Categories used for annotating the harvested trend keywords.

is merged with the other *Internet memes* category, it would have the same percentage of trends as sports news.

An interesting entry is the presence of changes to the Twitter official web service and/or API, which accounted for 6 trends (1% of the total) — this indicates that users have a meta-discussion about Twitter itself in the context of everyday Twitter messages[3].

As an aside, I also calculated the amount of spam topics (cf. Cheong and Lee [2009]) in this survey. This amount turns out to be just about 1% (6 trend strings) of all total strings.

8.1.4 Goal 3: Trend Localization

Based on the explanation by *What The Trend?*, I am also able to pinpoint a particular Trending Topic to a country or region, as some trends are localized and/or location information is provided from the crowd-sourced explanations. This is documented in Experiment 8.3.

Experiment 8.3. *Localizing the country-specific trends found in the Trending-Topics-Nov2009 dataset.*

Method: By checking the crowd-sourced *What the Trend?*-supplied definition or Google search results, one can deduce country-specific Trending Topics by various means; e.g. use of a different language, mentions of geographic areas (cf. [Humphreys et al., 2010]), or region-specific proper names. I inspect all 466 records in the *Trending-Topics-Nov2009* dataset for such cues, and associate each Trending Topic to a certain country if possible. Trends which are of global concern or where the country or region is not explicitly specified are not annotated with a specific country.

Results and Discussion: Out of the 466 total records, 253 trends have been associated with a particular country; with the remaining 213 not considered for analysis due to lack of localization cues. Some trends are associated with two or more countries; these are tagged with more than one country label. Hence, absolute percentages are not used due to this potential overlap (i.e. the sum of percentages might exceed 100%).

The table in Figure 8.4 contains the breakdown by country, while the bar graph illustrates the trend statistics aggregated by region (regions are adapted from the grouping proposed in [Java et al., 2009] and [Krishnamurthy et al., 2008]).

Looking at the aggregated results of trends by geographic region, the distribution of trends approximate the distribution of global Twitter users, as shown in studies by Java et al. [2009] and Krishnamurthy et al. [2008], albeit with minor differences of the last two regions' rankings.

From Experiment 8.3, I was able to infer that the Twitter user base is predominantly focused in the United States where Twitter had its origins, which explains why certain trends are highly specific (localized) to the US. The majority of sporting events surveyed

[3] This may be beneficial to research on social presence and similar fields: examples of related research include Mischaud [2007] and McNely [2009]

Country	Trends
US	149
UK	49
Brazil	10
Indonesia	10
Phillipines	8
France	6
Ireland	6
Canada	5
Australia	3
Korea	3
New Zealand	3
Japan	2
Belgium	1
Chile	1
Mexico	1
Netherlands	1
Puerto Rico	1
Singapore	1

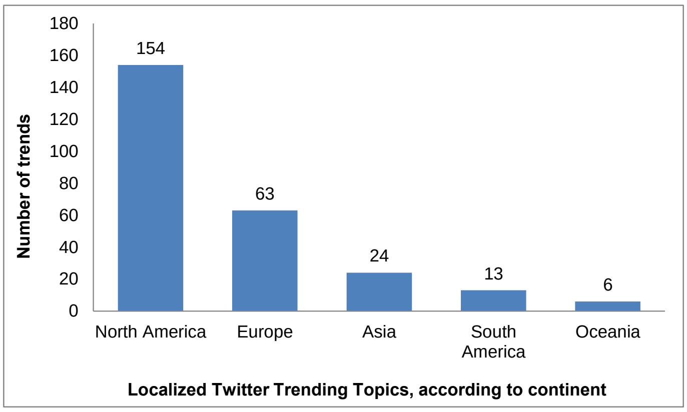


Figure 8.4: Table (on the left) shows the breakdown of trends associated with a particular country. The bar graph (on the right) shows the aggregated trends by region.

involve American-based sports — such as *National Football League*, the *National Basketball Association* tournament and *Major League Baseball* — all of which are highly popular in the US. Entertainment news such as those pertaining to American celebrities and artists are also commonly found in the *Trending-Topics-Nov2009* dataset.

8.1.5 Goal 4: Statistics on Trending Behavior

In the final goal of this case study, I analyze the popularity of the topics in the *Trending-Topics-Nov2009* dataset. This is done by monitoring their duration on the top ten Trending Topics list. As explained in Goal 1 (Section 8.1.2), a counter is used to track how many intervals a particular trend stays on the list of top ten trends, and also a record of the highest possible rank a topic has on that list.

Experiment 8.4. *Analyzing the temporal persistence of Trending Topics in the Trending-Topics-Nov2009 dataset.*

Method: Recall that in Goal 1 (Section 8.1.2), each of the trends in the *Trending-Topics-Nov2009* dataset are saved together with information about how long a trend persists in the Trending Topics list, along with statistics on the minimum and maximum ranking it has achieved.

For each trend, I extracted its highest rank and the number of 5-minute time intervals recorded. Pairs of the two variables for each trend in the *Trending-Topics-Nov2009* dataset are then used to generate a scatter-plot. I then attempt to fit a model to the scatterplot data, and determine its fit using the R^2 coefficient. The R^2 is simply the square of the

Pearson product-moment correlation coefficient [Pearson, 1895], defined earlier in Equation 5.2.

Also, to visualize the kind of trends with the most 'staying power' on the Trending Topics list, I plot the temporal distribution of all the 466 trends in *Trending-Topics-Nov2009* with respect to the time it spent as a Trending Topic.

Results and Discussion: Firstly, Figure 8.5 is the scatterplot illustrating the relationshop between highest rank (X, on the horizontal axis) and total time as a Trending Topic in terms of 5-minute intervals (Y, on the vertical axis) for each of the 466 trends.

By fitting the plot to an exponential model ($Y = 11.915e^{-0.406X}$), an approximate Pearson R^2 value of 0.60 is obtained (Equation 5.2).

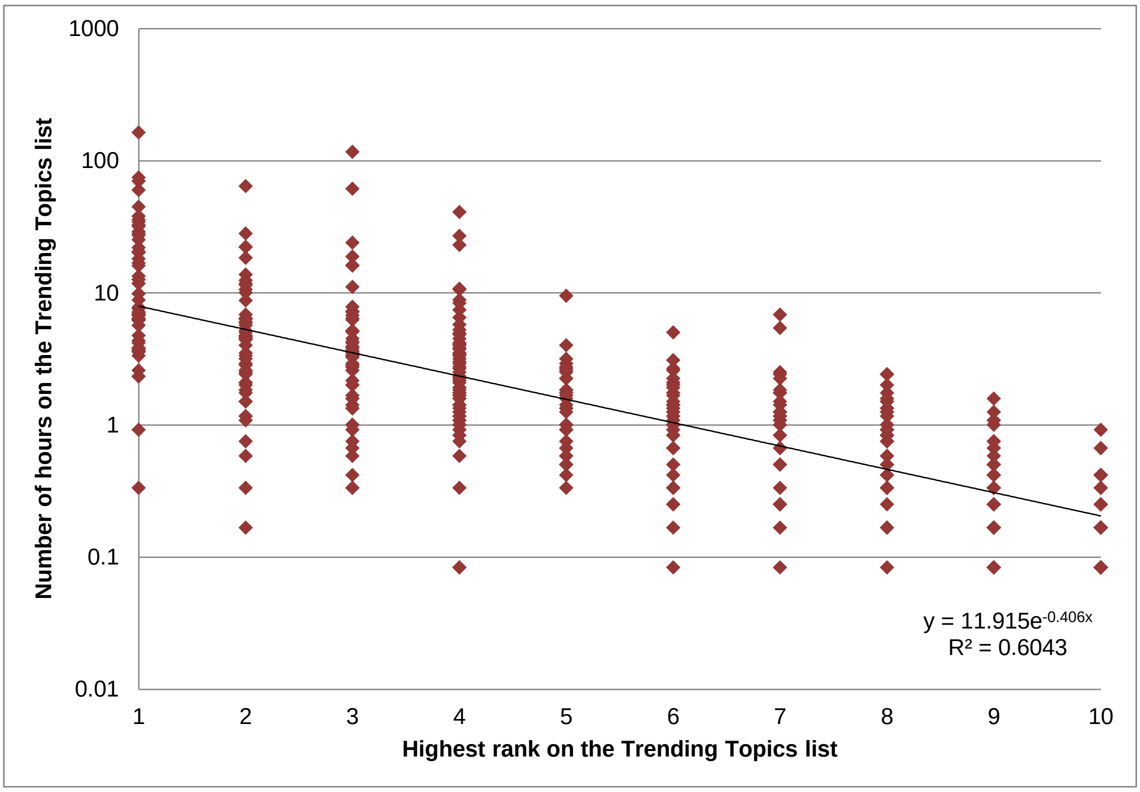


Figure 8.5: Trends' highest rank versus the number of time intervals recorded.

Secondly, Figure 8.6 illustrates the distribution of the 466 trends in *Trending-Topics-Nov2009*, with respect to the time it spent as a Trending Topic.

The following observations can be noted:

1. **Persistence of highly popular trends:** Trends which are persistently highly-ranked spend a long time in the Trending Topics list; this indicates a continuous mention of the topic by Twitter users. However, there are some trends that suddenly peaked to the top five of the list but rather quickly 'died off' (i.e. fell off the Trending Topics list) due to lack of attention.

   These two patterns (Figure 8.5) are evident in the difference of spread of '*trending time*', or the total amount of time a topic is listed in Trending Topics. The different spread of *trending time* is evident between the most popular trend (trend rank $X = 1$:

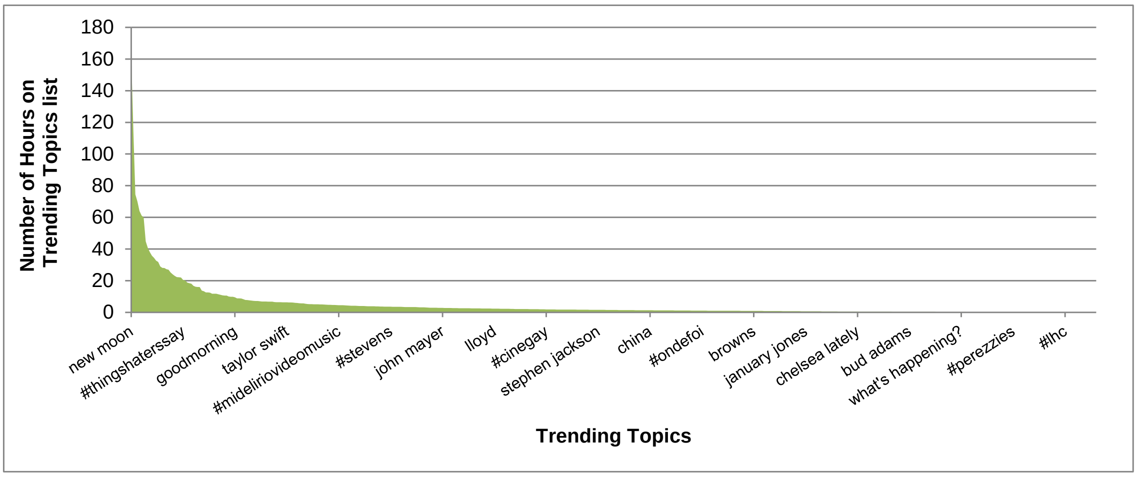


Figure 8.6: Distribution of trends on the Trending Topics list versus total time they were trending for.

20 minutes—163 hours inclusive) and the least popular trend (trend rank $x = 10$: only 5—55 minutes inclusive). Several causes behind this phenomenon have been identified:

- This may indicate the usage of alternate strings — such as `#hashtags` or different variations of the topic phrase — by users to carry on the conversation. Example: mentions of the celebrity Oprah Winfrey, where the phrases `#oprah` (in `#hashtag` notation) and `oprah` (without hashtag notation) are used interchangeably.
- Another cause for this is the existence of Twitter spam epidemics that quickly fell off due to user awareness and preventative measures by Twitter (cf. observation in [Cheong and Lee, 2009]). Example: the phrase `and lots more` found in spam tweets, which was trending for a brief period of time before being relegated from the Trending Topics list.

2. **Persistence of less popular trends:** Most of the trends which persist only for a brief period of time (cf. the trends in Figure 8.5 with a low *trending time*) tend to drop below the top ten rankings immediately after it 'trended'.

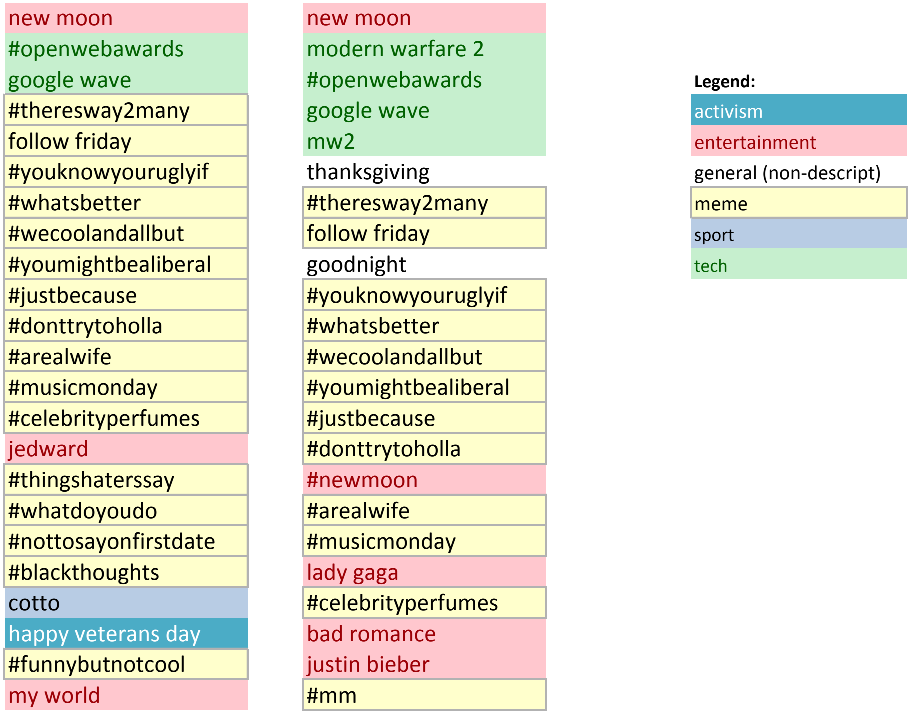
3. **Total trending time:** The top 5% of trends — the first 26 trends out of a total of 466 — persisted for over half the total observation time (Figure 8.6); the total observation time is defined as the total *trending time* by all trends in the *Trending-Topics-Nov2009.* In fact, the distribution of each topic's *trending time* can be modeled with a negative exponential distribution, similar to the distribution of source strings and domains in URL strings (cf. Sections 5.4.3 and 5.5.2).
In Experiment 8.4, the curve equation $Y = 116.52^{X/-12.373}$ fits the given data, with Pearson $R^2 = 0.8855$ via Equation 5.2. (X = ranking of a given trend in terms of total trending time; and Y = observed frequency of said trend cf. vertical axis in Figure 8.6).

The final experiment to wrap up Goal 4 is observing the composition of the top 5% of the trends, as detailed in Experiment 8.5.

Experiment 8.5. *Investigating the category distribution found in the Top 5% of the trends in Trending-Topics-Nov2009; in terms of: (a) time as a Trending Topic; and (b) highest rank on the Trending Topics list.*

Method: By sorting the collected *Trending-Topics-Nov2009* dataset from Goal 1 (Section 8.1.2 in descending order of time, I filtered out the top 26 tweets (i.e. top 5%) and colored them according to category. This is repeated, albeit with the sort key being highest rank on the Trending Topics list.

Results and Discussion: The top 5% of tweets, resulting from both sort orders, color-coded according to category, is illustrated in Figure 8.7.



new moon	new moon
#openwebawards	modern warfare 2
google wave	#openwebawards
#theresway2many	google wave
follow friday	mw2
#youknowyouruglyif	thanksgiving
#whatsbetter	#theresway2many
#wecoolandallbut	follow friday
#youmightbealiberal	goodnight
#justbecause	#youknowyouruglyif
#donttrytoholla	#whatsbetter
#arealwife	#wecoolandallbut
#musicmonday	#youmightbealiberal
#celebrityperfumes	#justbecause
jedward	#donttrytoholla
#thingshaterssay	#newmoon
#whatdoyoudo	#arealwife
#nottosayonfirstdate	#musicmonday
#blackthoughts	lady gaga
cotto	#celebrityperfumes
happy veterans day	bad romance
#funnybutnotcool	justin bieber
my world	#mm

Legend:

activism
entertainment
general (non-descript)
meme
sport
tech

Figure 8.7: Top 26 strings, in order of (a) time on the Trending list, and (b) highest rank on the Trending list.

Using either sorting criteria, observe that most of the top trends are on memes originating from Twitter or other parts of the Internet (excluding those started by celebrities). Technology-related news and entertainment news are also visible in the list. However, news on activism and sport rarely reach the top ten Trending list.

8.1.6 Case Study Summary and Discussion

The case study in this section has shown how an exploratory survey of the top trend strings appearing on the Twitter Trends list can reveal much information about the interests and topics of discussion among the *Twitterverse*.

To a certain extent, latent emergent behavior can be observed simply by assigning meaning and context to the trends themselves, as have been done here. By also tracking the movement in rank of trends over time, I was able to oversee how topics populate the Trending Topics chart with respect to its popularity; with applications in studying memetic behavior and viral information spread.

From this study, several areas can be further explored in the future. By combining a 'big-picture' analysis of trends with the inferences found from user and message metadata (Chapter 4), one can study the user base contributing to such topics in much detail. This allows for an observation as to how the user base's demographics and usage habits is reflected in the context of Trending Topics. Memetic behavior can also be tracked on Twitter by coupling the observations found in this case study with, say, analysis of the Twitter social graph and user inferences.

8.2 Measuring Trend Persistence using the On-Demand API

In a preliminary study conducted in 2009 during the start of my PhD (Section 6.2), I tinkered with the idea of analyzing temporal 'spikes' and persistence of trends based on the frequency of messages extracted from the `search` API.

To quickly recap, in Cheong and Lee [2009], I have first created the *Twitter-SWSM2009 dataset*. This dataset consists of 7,215 tweets from six different topics, and three different temporal patterns. Parts of Sections 4.6, 4.7, 4.8 have been dedicated to discuss my preliminary studies of metadata inference in Cheong and Lee [2009]. Sunsequently, Section 6.2 has discussed about self-organizing map clustering on the inferences resulting from the use of the *Twitter-SWSM2009 dataset*.

In this section (Section 8.2), I shall now document my preliminary analysis and categorization of temporal spiking/trending behavior exhibited by tweets, as published in [Cheong and Lee, 2009]. Despite not being the primary focus of this thesis, I find it appropriate that my case study findings from Cheong and Lee [2009] be documented here.

Reviewing the literature published post-2009, Kwak et al. [2010] have independently come up with similar analyses of temporal behavior of tweets (reviewed in Section 3.3.2), approximately a year after my analyses were published [Cheong and Lee, 2009]. This vindicates my original study of spiking behavior of tweets and their temporal persistence.

8.2.1 Study Preliminaries

Usage of UIDs as Measure of Time

One of the novelties of my study [Cheong and Lee, 2009] is that I use the unique message identifier (UID) generated by Twitter for each message (Section 4.3.1) instead of the more common usage of the tweets' datestamp/timestamp directly to measure time (x-axis). As far as I know, no prior research before [Cheong and Lee, 2009] has used the UID as a measure of time.

The benefits of using the UID (instead of time) are its relative ease of use, and the frequency of UID generation over time is more or less stable. Should the frequency of generation of new UIDs vary in the future (e.g. due to increased popularity of Twitter; or conversely, declining usage), the UID frequency can be a reliable indicator for future trend/spike analysis.

The UID frequency, determined by dividing the UID interval (between first and last messages for each of the 6 case studies) with the date range (between the first and last tweets), is on average 111 UIDs per second, which is a good reflector of the current rate of message flow on Twitter [Cheong and Lee, 2009].

Study Design

Section 6.2 has earlier highlighted how my work in [Cheong and Lee, 2009] has resulted in the creation of the *Twitter-SWSM2009 dataset* for metadata inference and clustering. The complete listing of the topics are as per Table 6.1 in Section 6.2.1.

I conducted my analyses on the entire set of 7,215 tweets containing trend- and non-trend-related keywords, found in the *Twitter-SWSM2009 dataset*. The frequency distribution of tweet count versus time (expressed as unit UIDs) was created for tweets in each topic. As the `search` API rate-limits the amount of tweets for a particular topic (refer Section 4.1.3), this effectively time-slices my observations, allowing for the observation of temporal behavior of tweets from one of two perspectives:

1. **Temporal persistence**, in terms of the maximum UID range (i.e. the period of time) in which tweets regarding a particular topic backdates to; given a constant window of tweets. The fixed window of tweets is a direct consequence of using the Twitter `search` API with a hard rate limit of 1,500 tweets (Section 4.1.3). Within a sample of tweets, a topic with long temporal persistence would exhibit a longer UID range, and vice-versa.

2. **Spiking behavior**, in terms of a sudden 'spike' of frequency of tweets per unit time. This is based on spike analysis on trends from existing phenomena such as the stock market [Choudhury et al., 2008] and blog posts [Fukuhara et al., 2005].

From that, I could identify three broad temporal persistence patterns of topics for both trending or 'spiking' [Gruhl et al., 2004] topics and non-trending topics:

1. **Long-term trend persistence:** Sparsely discussed due to obscurity. Any spike in the messaging frequency will decay relatively quickly. Topics like this can be accessed by the Twitter Search API up to approximately 20 days (a soft constraint), but rarely exceed the maximum retrieval results of 1,500 tweets [Cheong and Lee, 2009].

2. **Medium-term trend persistence:** Topics which are either generic terms which are commonly talked about but do not warrant a high number of tweets; or sustained 'trailing patterns' [Fukuhara et al., 2005]. The trailing patterns in the latter case [Fukuhara et al., 2005] are due to an existing spike that occurs beyond the REST API-imposed 1500 tweet boundary, but the discussion on the topic is trailing as the

dataset is harvested [Cheong and Lee, 2009]. Such topics can range for a period from half a day to approximately a few days [Cheong and Lee, 2009].

3. **Short-term trend persistence:** Such topics are high-volume in nature and can be a very commonly-talked about term which does not exhibit spiking behavior. These can also be high-volume topics captured using the Twitter API in the middle of a spike. Topics like these, during the moment of REST API data capture, will only backdate up to a few hours when accessed. Topics as these can be categorized as those in the 'graduated increase pattern or in the middle of a 'periodic pattern [Fukuhara et al., 2005; Cheong and Lee, 2009].

With respect to the above list, the following subsections document examples and the observable properties of such topics.

8.2.2 Case 1: Long-term Persistence (ranging from few hours to days)

Topics with long-term trend persistence, or simply '*long-term topics*', have a relatively low occurrence in the Twitter public timeline overall.

The control term `Revolverheld`, a German alternative-rock band, was used to study the trend of an obscure, non-trending topic, which exhibits no spiking behavior whatsoever. Figure 8.8 refers to the pattern captured by the obscure topic `Revolverheld`, over a period of approximately 20 days.

The obscurity of the topic `Revolverheld` is seen by the fact that the frequency of mentions of this topic remain at a minimum level: three tweets or lower per 770 thousand UIDs (over a period of ~2 hours). In fact, the Twitter chatter for such an obscure topic can be regarded as a Poisson process; with a maximum likelihood estimate of $\lambda = 0.2360$.

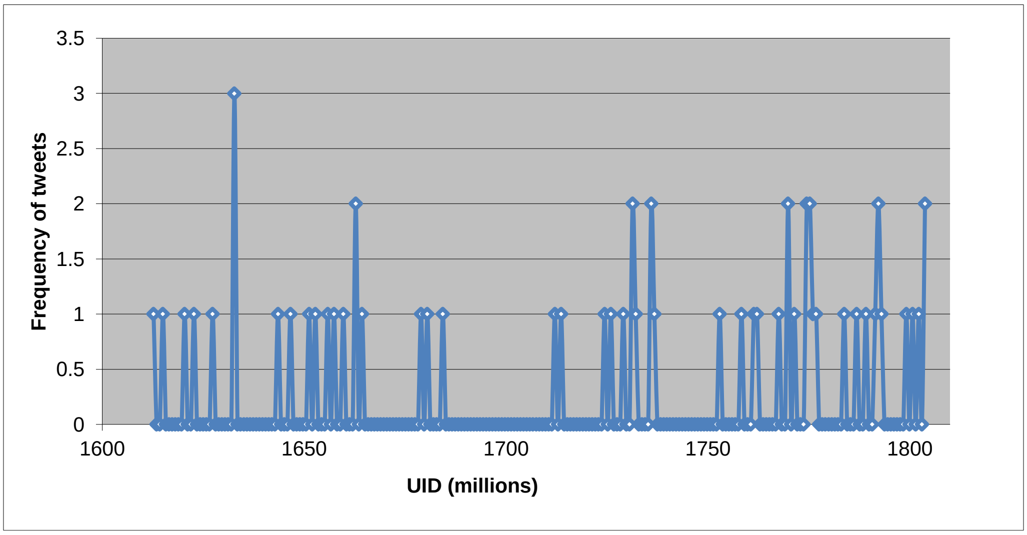


Figure 8.8: UID (in millions) versus tweet frequency plots for long-term topic keyword `Revolverheld`, an obscure, non-trending topic. Note the magnitude of the y-axis which is in the order of ones.

The Twitter Trending Topic `Nizar` (a Malaysian politician involved in a constitutional crisis) was chosen to study the trend of a *quick spike*. Quick spikes are also known in existing literature as:

- The Slashdot effect [Adler, 1999], named after the Slashdot website where featured articles which gain a spike in popularity will be inundated with web traffic and exhibit spiking behavior.
- A *sensitive pattern*, as per Fukuhara et al. [2005], who studied similar spikes in blog articles and real world temporal data.

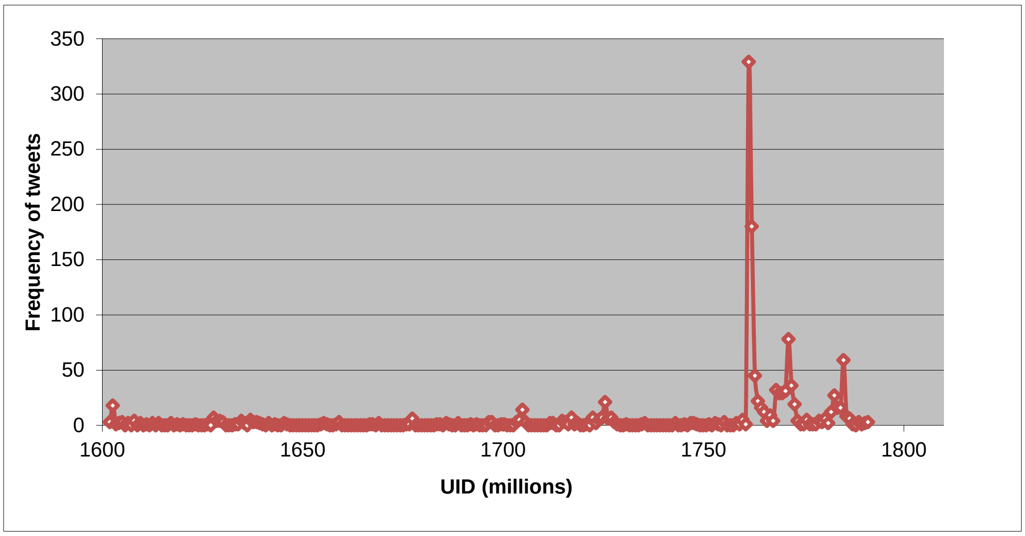


Figure 8.9: UID (in millions) versus tweet frequency plots for long-term topic keyword `Nizar`, a quick-spiking topic. Note the magnitude of the y-axis which is in the order of hundreds; contrast this with Figure 8.8.

Figure 8.9 illustrates the temporal behavior of `Nizar`. This topic keyword spiked at 11 May 2009 (at approximately 07:00 GMT), which directly corresponds to the real-world event of the Malaysian courts passing judgment on Mr. Nizar's political case [Mageswari and Goh, 2009].

The main spike consisted of 329 tweets in a window of approximately two hours, which was followed by two spikes of lesser magnitude (78 and 59 tweets respectively). Over the observation period, `Nizar` to third place on the global Twitter Trending Topics list before gradually fading off, corroborating the sensitive decay pattern discovered in Fukuhara et al. [2005]. In fact, the work by Kwak et al. [2010] — published a year after my work in this chapter [Cheong and Lee, 2009] — corroborated my findings; such news topics exhibit similar spiking behavior and are termed "exogenous critical" [Kwak et al., 2010].

8.2.3 Case 2: Medium-term Trend Persistence (few hours)

Topics with a medium window of persistence have a significantly shorter range of UIDs (and in relation, time), compared to long-term topics. Such 'medium-term topics', when fetched from the Twitter `search` API, are accessible as far back as the `search` API's retrieval limit of 1,500 tweets.

Figure 8.10 illustrates the tweet frequency over time (in terms of UIDs) of two such '*medium-term topics*':

1. `H1N1`, the subtype of influenza responsible for the 2009 Swine Flu pandemic.

2. `TwitHit`, a keyword found in spam tweets generated by Twitter users who inadvertently supplied their login credentials to a spamming site [Cashmore, 2009b].

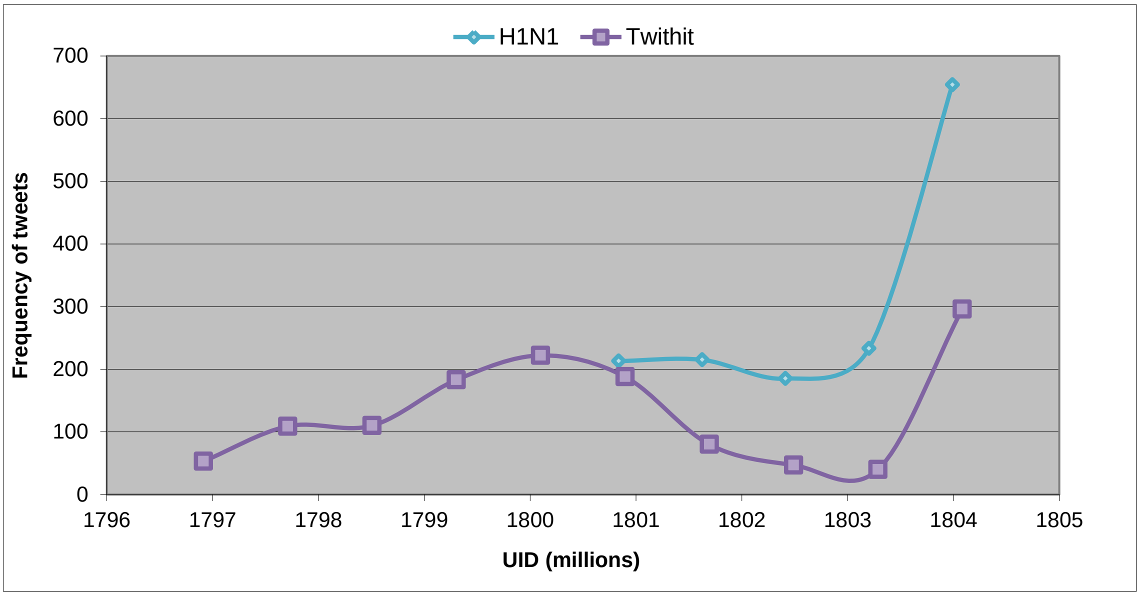


Figure 8.10: UID (in millions) versus tweet frequency plots for medium-term topic keywords `H1N1` (blue-colored series) and `TwitHit` (purple-colored series).

The blue-colored series in Figure 8.10 illustrates the temporal distribution of Twitter chatter about the H1N1 swine flu pandemic, which was a Trending Topic. The tweet data (as part of the *Twitter-SWSM2009 dataset*) was captured about halfway through the time H1N1 was on the Trending Topics list. This trend is classified as a trailing pattern [Fukuhara et al., 2005]: persistent Twitter chatter on this topic has caused it to be included in the Trending Topics list ever since the outbreak of H1N1 began in early May 2009.

Such a medium-term topic has a range of 3.1 million UIDs (compared with the short-term topics with range of 200 million UIDs). It can be observed that medium-term topics generate hundreds of mentions within a histogram interval (as defined in Section 8.2.2: 770 thousand UIDs corresponding to approximately 2 hours). The observed frequency corroborates with the definition of a *spike*, in the case of the `Nizar` keyword in Section 8.2.2.

Another medium-term topic that exhibits spikes is the case of the keyword `TwitHit` (purple-colored series in Figure 8.10) which originated from spamming activity [Cashmore, 2009b]. This trend can be roughly categorized as a *sleeper hit* [Fukuhara et al., 2005]: a topic which rose in popularity from relative obscurity to a trending topic. The sleeper hit pattern is discovered in many other Twitter trends, e.g. similar Twitter scams or an unexpected catastrophe of great media significance.

The time range for `TwitHit` is approximately 8 million UIDs (nearly triple the one from `H1N1`). However, based on the same histogram plot interval (approximately 770 thousand UIDs ≡ 2 hours), the `TwitHit` topic still fits the category of a medium-term trend.

As I was able to capture data from the first occurrence of the `TwitHit` spam, I was also able to accurately investigate the following two interesting curiosities regarding spam on Twitter. To my knowledge, this was the first time such a study on Twitter spam epidemics was conducted and published [Cheong and Lee, 2009]. Two properties were discovered from my study:

- **The exact time in which spiking behavior is exhibited by spam:** From Figure 8.10, one can easily see that the spike or 'trending' characteristic of the keyword `TwitHit` begins at the 9th interval. From the data, the spike occurred at 6.37 million UIDs, i.e. approximately 16.55 hours from the start of the outbreak.

- **Message generation in a spam epidemic:** By introspection of the raw `TwitHit` message data captured in the *Twitter-SWSM2009 dataset*, I observe that the occurrences of `@user`-based reply tweets do not occur until near the end of the trend. The first increase in messaging activity is due to automated spamming programs generating `TwitHit` spam tweets. This is followed by an increase in `@user`-based tweets at the end of the survey period, consisting of users discussing their experiences as victims of the spam.

8.2.4 Case 3: Short-term Trend Persistence (less than an hour)

Finally, certain topics exhibit short-term trend persistence; these are simply known as 'short-term topics'.

When such short-term topics are retrieved using the `search` API, the maximum set of 1,500 tweets returned from the Twitter API fits within a period of 4–5 hours. This is the result of a disproportionately large amount of tweets generated by users. Such short-term topics are usually related to the high visibility of top-ranking Trending Topics on Twitter at any given moment.

Figure 8.10 illustrates the tweet frequency over time (in terms of UIDs) of two examples of '*short-term topics*':

1. `Grey's Anatomy`, referring to the name of a US television drama series, which aired its fifth season finale in 2009 on primetime television. This phrase was on the Trending Topics list at time of creation of my *Twitter-SWSM2009 dataset* [Cheong and Lee, 2009].

2. `Coffee`, a control term (non-Trending Topic), referring to the everyday beverage.

The high volume of Twitter chatter succeeding the season finale of the `Grey's Anatomy` drama series (the green-colored series in Figure 8.11) caused this topic to have the smallest relative range of UIDs; which is characteristic of short-term topics.

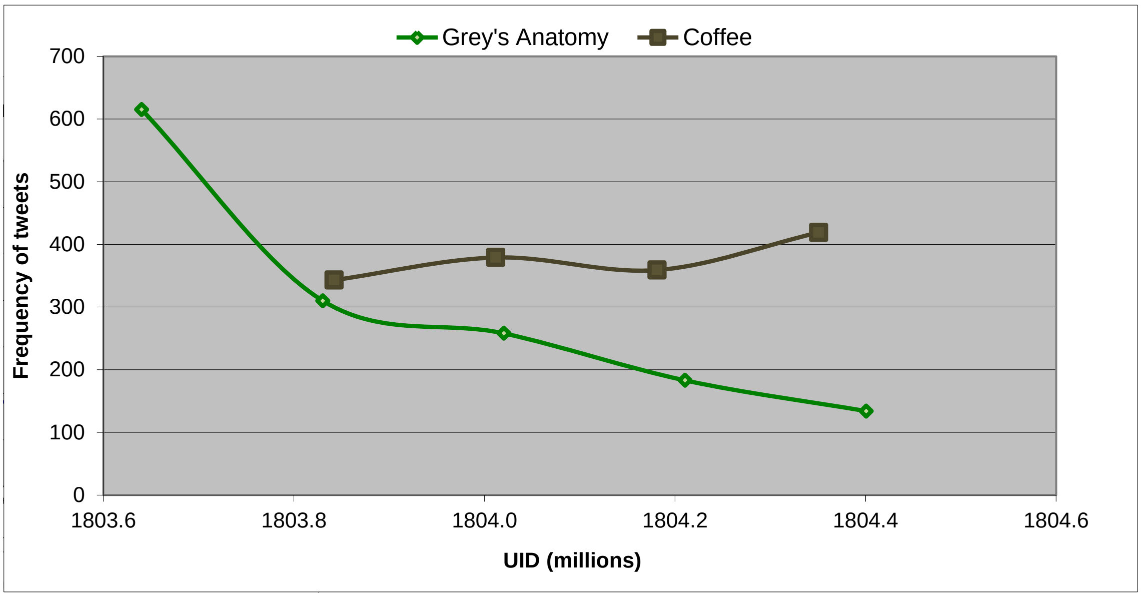


Figure 8.11: UID (in millions) versus tweet frequency plots for short-term topic keywords for `Grey’s Anatomy` (green-colored series) and `coffee` (brown-colored series).

I started capturing tweet data for this trend as soon as it jumped to first position on the Twitter Trending Topics list. I observed that the maximum set of results from the `search` API for `Grey’s Anatomy` topic encompasses a time period of roughly 950,578 UID units, equivalent to approximately 2.4 hours. For clarity, the histogram interval used in Figure 8.11 is approximately 190 thousand UIDs, or roughly 30 minutes, due to the volume of tweets in such a short time window. Note that the voluminous number of tweets make it impossible to go beyond the 1,500 message limitation of the `search` API. As such, the available range of UIDs, and correspondingly the observation window, is rather limited.

It is important to note that the keyword `coffee`, though exhibiting behavior of a long-term topic, did not show up in the Trending Topics list. This is because `coffee` is a relatively common term in everyday vocabulary. From empirical observation of the Trending Topics list, such everyday terms are not considered as a Trending Topic [Cheong and Lee, 2009]. Similar filtering methods are used in spike detection studies where common words are excluded as their uses in everyday chatter are frequent, and that such words are not proper nouns [Gruhl et al., 2004].

8.2.5 Discussion and Study Conclusions

The presence of retweet (`RT`) messages are common in almost all of the Trending Topics in this study [Cheong and Lee, 2009], which in turn contributes to the overall chatter of a trend. This is not unlike how email forwarding and blog linking behave in contributing to the memetic spread of a topic, cf. [Smith et al., 2005; Arbesman, 2004]. Replies are predominantly found on topics with high user interaction (with the exception of the *TwitHit* case, as will be discussed later). The interesting part is the prevalence of the keyword Trend only on topics which already have been included in the Trending Topics list: messages such as these are usually self-referential (to the trend keyword itself), or piggybacking on the term to generate more views; all of which are typically tactics employed by aggressive marketing campaigns and spammers.

8.3 Concluding Notes

Chapter 8 has provided a coverage of two studies, which resulted as secondary investigations in my research for this thesis. Despite the fact that these eclectic approaches are not the main foci of the thesis, such studies have nonetheless provided insight into trends on Twitter, with respect to the 'life' of Trending Topics, and the persistence and trending characteristics of popular topics.

The following chapter shall conclude my thesis, in which I will recap what has been achieved so far in accomplishing the subgoals set forth in the Introduction (Chapter 1), and ultimately leading to the answering of my main research question in this thesis.

Chapter 9

Conclusion

"*Well I, believe, it all, is coming to an end,*
Oh well, I guess, we're gonna pretend,
Let's see how far we've come..."

— Matchbox Twenty,
How Far We've Come (2007).

To conclude this thesis, I will sum up my work in the preceding chapters. In this conclusion, I will demonstrate how my contributions serve to answer the central underlying question of my thesis: *how can we discover new knowledge from a combination of the user and message domains on a microbloging service such as Twitter, that would ultimately lead to a better understanding of real-world behavior and events?* In the process of answering this central question, I have achieved several subgoals as set out earlier in the introduction to this thesis (Chapter 1).

First, I described the underlying principles behind Twitter, a microblogging service; how users make use of the service; and idiosyncracies resulting from its use (Chapter 2). Then, I examined the state-of-the-art research with respect to Twitter and related technologies. This examination revealed that extant work on combining both Twitter's user and message domains is limited, and that there is a gap in existing knowledge — the disparity in research that encompass both domains — that needed to be filled (Chapter 3).

In Chapter 4, I identified the quality and quantity of metadata that could be harvested from Twitter, and evaluated the interfaces available for doing so. Armed with these metadata, I was able to build novel algorithms and metrics to generate inferences on real-world demographics, online presence habits, and tweeting/communication patterns, that seek to make sense of the raw metadata omnipresent on Twitter. Inspired by existing literature in fields ranging from computer–human interaction to sociology, I was able to augment my algorithms to scale well in terms of speed and performance when applied to very large datasets (of the magnitude of millions). I was also able to show that my proposed metrics and algorithms are able to detect anomalous features present in the Twitter user base.

Subsequently, in Chapter 5, I designed two frameworks to automate the process of metadata discovery, collection, and archiving, using the two variants of the Twitter API.

Using these frameworks, I captured a large dataset of real-world Twitter activity spanning seven days, culminating in millions of user and message metadata records. With my discoveries and contributions from the prior Chapter 4, I was able to conduct a large-scale analysis of Twitter using the gathered data, and was able to construct a world-view of Twitter as it currently stands.

Given the results of the application of my methods on raw metadata, there is a need to find commonalities and hidden patterns to make more sense of this new-found knowledge. Chapter 6 introduced the approach of clustering the results of metadata-based inferences for the user and message domains. I studied the current literature for a suitable clustering algorithm given our characteristic data, and have experimented successfully with the Kohonen self-organizing map technique. I then conducted testing to assess the suitability of SOM to cluster and visualize common users (and their messaging behavior), given collections of tweet metadata as input.

The approaches in the preceding chapters needed to be tested on real-world situations; which is why I performed three case studies to test the efficacy of my contributions presented thus far. Among these is, foremost, a longitudinal study of the 2009–2012 Earth Hour campaigns to construct and analyze patterns of user activity (with a comparison with the real-world implications of the campaign). The result of this study has proven that Twitter time-series data can be effectively used in tandem with real-world statistics and figures to determine a link between real-world events and Twitter activity (in this case, real-world electronic activism campaigns). The following case study is where I produce a framework to utilize the knowledge and raw metadata from Twitter as a potential terrorism informatics platform; which illustrated the potential usefulness of such a framework during terrorism scenarios, and the wealth and quality of data at the disposal of stakeholders in such events (e.g. law enforcement, the military, and researchers). The last case study uncovers the behavior of users and their corresponding tweeting strategies, vis-à-vis the 2011 London Riots as they unfolded on the Twitterverse. This last case study — dealing with riot and chaotic events — resulted in the unearthing of useful emergent properties from Twitter metadata, that can be used to study the symptoms and evolution of the actual rioting event.

Finally in this thesis, I presented an assortment of analytic techniques and trend studies on Twitter obtained throughout the research conducted for this thesis that do not fit into prior categories, but nonetheless can provide a useful insight on the idiosyncratic behavior peculiar to Twitter, and a foundation for future related work (Chapter 8).

Figure 9.1 summarizes my contributions in terms of the successful completion of the five sub-goals set forth in the thesis, and how all of my contributions so far have led to the answering of the research question underlying this thesis.

Despite conducting my research contributions as rigorously as possible, several limitations were inevitable. Perhaps the biggest limitation encountered and documented throughout (e.g. Section 4.2.3) is the lack of a complete set of Twitter metadata for thorough analysis. As reiterated especially in Chapter 4, it is at best currently infeasible (at worst, impossible) to obtain a complete set of metadata for a given timeframe or a given

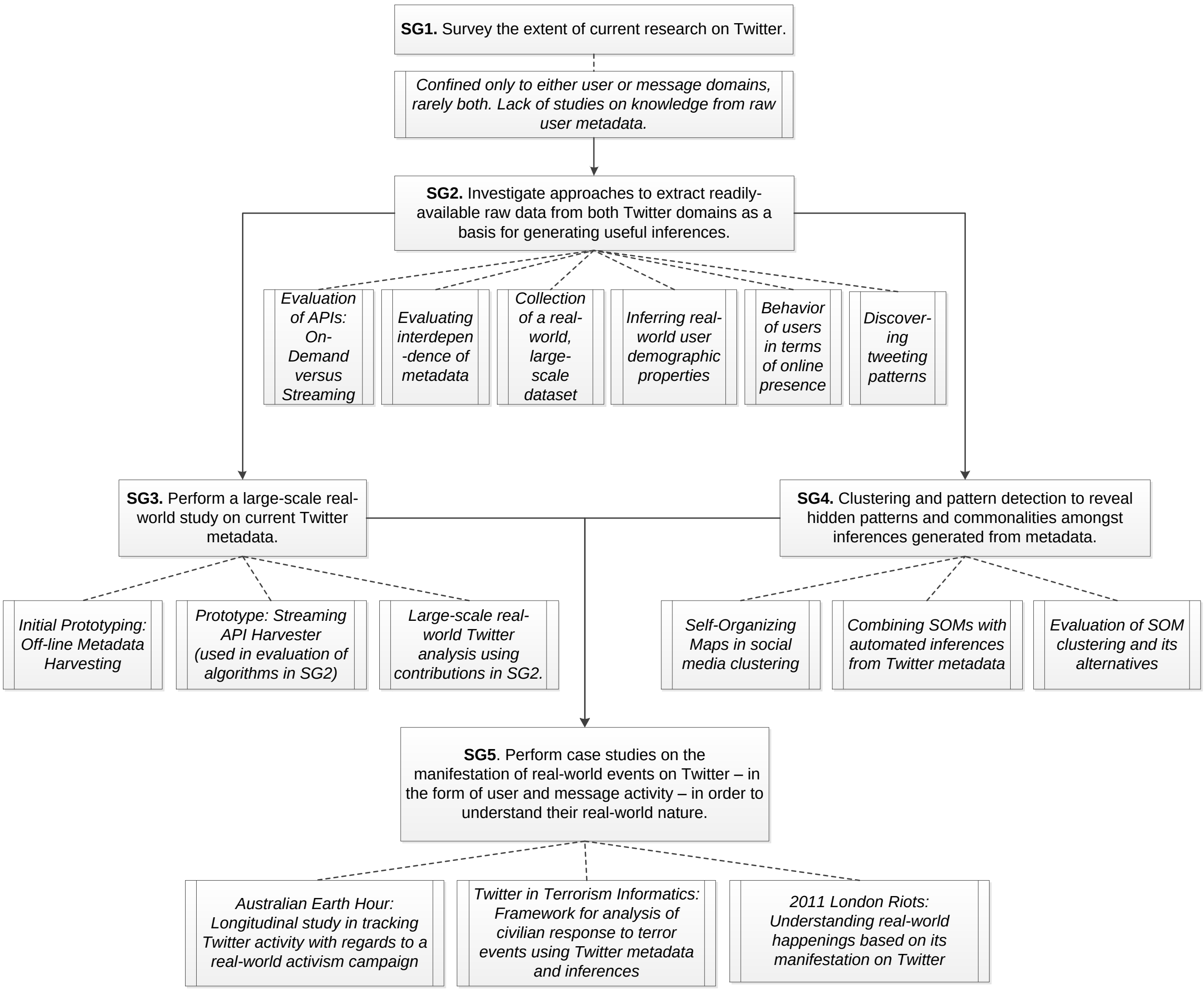

SG1. Survey the extent of current research on Twitter.
Confined only to either user or message domains, rarely both. Lack of studies on knowledge from raw user metadata.
SG2. Investigate approaches to extract readily-available raw data from both Twitter domains as a basis for generating useful inferences.
Evaluation of APIs: On-Demand versus Streaming
Evaluating interdepen-dence of metadata
Collection of a real-world, large-scale dataset
Inferring real-world user demographic properties
Behavior of users in terms of online presence
Discover-ing tweeting patterns
SG3. Perform a large-scale real-world study on current Twitter metadata.
Initial Prototyping: Off-line Metadata Harvesting
Prototype: Streaming API Harvester (used in evaluation of algorithms in SG2)
Large-scale real-world Twitter analysis using contributions in SG2.
SG4. Clustering and pattern detection to reveal hidden patterns and commonalities amongst inferences generated from metadata.
Self-Organizing Maps in social media clustering
Combining SOMs with automated inferences from Twitter metadata
Evaluation of SOM clustering and its alternatives
SG5. Perform case studies on the manifestation of real-world events on Twitter – in the form of user and message activity – in order to understand their real-world nature.
Australian Earth Hour: Longitudinal study in tracking Twitter activity with regards to a real-world activism campaign
Twitter in Terrorism Informatics: Framework for analysis of civilian response to terror events using Twitter metadata and inferences
2011 London Riots: Understanding real-world happenings based on its manifestation on Twitter

Figure 9.1: Big-picture overview of the five subgoals in this thesis: each of them accomplished by my findings and contributions within each subgoal, listed in italics.

search query. One is able to obtain a sampled subset only; hence illustrating the need for a framework of data sampling and archival, tailored to the idiosyncrasies of Twitter (Chapter 5). Some of the reasons for the restrictive nature of Twitter data collection are cost, lack of resources on the part of Twitter Inc., and privacy/legal implications. A related limitation is the relative difficulty capturing metadata pertaining to important real-world events. To catch a significant event as it unfolds on Twitter requires one to continuously capture real-time metadata from Twitter in its entirety, which is taxing in terms of computing and network resources. The alternative to this is to search for backdated or archived metadata from Twitter, which is again infeasible, as Twitter Inc. severely limits the retrieval of archival data due to internal constraints.

Despite the shortcomings above, I have shown that my contributions within this thesis will lead to a better understanding, in terms of microblogging, who social media consumers are, how they communicate online, and how these patterns 'mirror' the real-world. My thesis has dealt with studying core metadata from both user and message domains, learning about their emergent features and how such features relate to the real world, and detecting the hidden patterns and commonalities found within the metadata. As a practical demonstration and test, I applied my new framework and algorithms to show that they can derive useful insights about real events, such as the 2011 London Riots. With this knowledge at hand, I was able to contribute to the further study of the people who form the heart of the Twitter microblogging service, and social media in general.

Also, as a consequence of work presented in this thesis, I have identified several potential areas for future research. This includes combining my novel inference metrics and algorithms with related real-world studies on human behaviour, scale-free networks, and technology usage habits; expanding my large-scale empirical study of the *Twitterverse* (at the same time further testing the scalability of my methods) with a larger collection of Twitter metadata than the one demonstrated in this thesis; and devising new methods of visualizing obtained inferences and patterns resulting from the use of clustering algorithms.

As a concluding note, my thesis has uncovered several broader implications in social media studies in general, with a specific focus on microblog research. These include applying similar methods on other social media (such as Facebook and Google+); near real-time interpretation of live Twitter metadata; and identification of emergent patterns with advanced clustering and heuristics. My research also paved the way for future work in dealing with ever-voluminous amounts of data in other contexts — such as an ensemble of methods combining filtering, aggregation, and standard analysis — that has been illustrated as part of this thesis.

Appendix A

Metadata on Twitter

For the sake of completeness, this appendix contains a technical overview of the various items of metadata found in user and message records harvested using the Twitter APIs.

A.1 Technical Descriptions of Metadata Fields

A.1.1 Message Metadata

Message text

The `text` field contains the raw 140-character-limited text in a tweet, which is visible to followers of its originating user. This text is mainly consumed by users of the Twitter front-end (e.g. website, mobile clients), and is the most visible attribute of a message. Substrings of this message could take the form of URLs, hashtags, `@user` references, and 'smileys' such as "`:)`".

Recent developments to the Twitter API have simplified the process of harvesting entities from the message text. Several examples include the presence of URLs, hashtags, and `@user` references, which are automatically summarized by Twitter in their own separate metadata fields when a tweet message is composed.

Message ID

`id` is a numerical variable containing the message ID that uniquely identifies a message internally in Twitter (end users rarely see it). Also, for ease of parsing, `id_str` is the same variable, formatted by the API as a string. The string equivalent is provided to "allow JavaScript and JSON parsers to read the IDs" [Twitter Inc., 2012a].

Software client/source

`source` contains a string identifying the software client used to publish this particular tweet. This can be one of several 'default' sources, or custom/third-party applications. The default or official sources are:

1. `web` refers to the official Twitter website, <`http://www.twitter.com/`>

2. `mobile` refers to the official Twitter mobile website, <`http://mobile.twitter.com/`>

3. `txt` refers to the official Twitter SMS service.

As for third-party applications, the `source` string will instead consist of the name of the application, as well as the URL of the application's website encapsulated with HTML anchor (<`a`>) tags.

Time-stamp

The `created_at` field is a time-stamp indicating when the tweet was composed.

It is of the format "`ddd MMM dd hh:mm:ss: zzzz yyyy`" where[1]:

- `ddd` is the abbreviated day of the week.
- `MMM` is the abbreviated month.
- `dd` is the two-digit day of the month.
- `hh:mm:ss` is the time expressed in two-digit hours, minutes, and seconds; with the colon as the delimiter.
- `zzzz` is the UTC offset.
- `yyyy` is the four-digit year.

An example of a formatted time-stamp would be "`Mon May 30 04:18:20 +0000 2011`".

In-reply-to

The metadata item `in_reply_to_status_id` (and its string equivalent `in_reply_to_status_id_str`) contains the ID of the parent message in which the current message is replying to. `in_reply_to_screen_name` and `in_reply_to_user_id` (and its string equivalent `in_reply_to_user_id_str`) on the other hand store the name and ID of the user to whom this message is directed to.

These items of metadata help application developers perform message threading or chaining, in which related replies or 'threads' are grouped together in a third-party application interface, to illustrate the relationship between messages.

Geo-tagging

This set of metadata items, new to Twitter (rolled out in 2010), is part of a Twitter API update to support geo-tagging and place identification for tweets [Twitter Inc., 2012a].

[1] Historical note: the format of this time-stamp has changed from "`ddd, dd MMM yyyy hh:mm:ss zzzz`" (e.g. "`Tue, 09 Jun 2009 09:37:07 +0000`"). The previous form of the message time-stamp is due to the inherent API differences between the Search API (which was one of the key sources of message data) and other APIs.

coordinates is a JSON object encapsulating the longitude and latitude of the message when it was composed (geo is a deprecated variable which is similar). place on the other hand is an object introduced by the Twitter API to meaningfully describe the exact place the message was composed; this includes information such as a geographic bounding box, place name, place type and country code.

As of time of writing, however, these metadata features are still in development (as per Section 4.1.4). Hence, for the purposes of this thesis, I mainly use the geographic information in the user domain as it has been readily available since the early days of Twitter and used in prior research, e.g. [Honeycutt and Herring, 2009; Longueville et al., 2009; Cheong and Lee, 2010c].

Authorship Summary

In the message metadata, the Boolean flag truncated indicates the truncation of a message when sent by SMS (i.e. exceeding 140 characters). This again is one of the improvements found in the evolution of the Twitter API. Meanwhile, contributors is another new piece of metadata that lists contributors who have authored this message; taken from the experimental Contributors feature [2] which allows multiple users to post messages as another user (i.e. by proxy).

Statistics on Retweets

retweeted is a Boolean flag indicating if the message has been retweeted by other users. If this was true, the number of retweets are specified by the retweet_count variable. However, if the amount of retweets for a message exceeds the order of a hundred, retweet_count will instead be the constant string 100+.

Entity Logging

Also, as part of improvements to the Twitter API, *entities*[3] related to a particular message are logged and made available as a JSON object in the message metadata. The user_mentions field stores a summary of other users who are mentioned in the current message with the @user convention: this includes the start and end position of the @user substring in the message body (indices); their user ID, screen name, as well as real name.

hashtags is an object containing hashtagged keywords mentioned in the current message with the #hashtag convention. As with user_mentions, the indices of each hashtag substring relative to the main message string are listed as well.

Finally, urls is similar to the above, but used to represent the presence of URL strings in the message text. The URL text (escaped with backslashes) and substring indices are listed, similar to user_mentions and hashtags. There are additional pieces of metadata unique to URLs, such as the expanded_url field which contains the full URL. display_url

[2] An API announcement, the "Developer Preview [of the] Contributor API" was written by Raffi Krikorian of Twitter Inc.: <http://groups.google.com/group/twitter-api-announce/browse_thread/thread/12273c2d03c1b606>

[3] *Tweet Entities*, as listed in the Twitter Developers documentation: <https://dev.twitter.com/docs/tweet-entities>

is a truncated version of the URL for display purposes (e.g. on the Twitter website) when the original URL in the message is too long and have to be truncated by Twitter (using their `t.co` URL shortening service) during the creation of a tweet.

The following hypothetical example illustrates the aforementioned `user_mentions`, `hashtags`, and `urls` entities, enumerated in JSON for a tweet containing the sample text "`@user123 G'day mate http://example.com #hello`".

```
"entities":{
    "user_mentions":
    [
        {
            "indices":[0, 8],
            "id":123456789,
            "id_str":"123456789",
            "screen_name":"@user123",
            "name":"John Doe"
        }
    ]

    "hashtags":
    [
        {
            "text":"hello",
            "indices":[40, 46]
        }
    ]

    "urls":
    [
        {
            "indices":[20, 38],
            "url":"http://example.com",
            "expanded_url":null,
            "display_url":null
        }
    ]
}
```

In late 2011, Twitter Inc. introduced the `media` entity, which encapsulates information about media linked to a tweet, such as photographs. Examples of fields in this entity include a unique ID (`id`), different versions of URLs linking to the given media file (`media_url`, `display_url`, `expanded_url`), and display dimensions (`sizes`). As the `media` entity is relatively new and still in development (as of time of writing, it only supports photos), a complete discussion is beyond the scope of this thesis.

A.1.2 User Metadata

User ID

Similar to its namesake in the message domain, `id` is a numerical variable containing the user ID that uniquely identifies this user internally in Twitter (end users rarely see it). Also, for ease of parsing, `id_str` is the same variable formatted as a string [Twitter Inc., 2012a], as with Message ID in Section 4.3.1.

Names

A Twitter user's real name, as supplied to Twitter during account creation is stored in the `name` field. `screen_name` on the other hand stores the screen name, or username, which is the name which is mainly utilized by end users in their communication with one another on Twitter: e.g. when addressing a user with the `@user` notation.

Description

The `description` field is a user-provided, 160-character limited text string that allows a user to describe himself with some brief profile information or bio.

Location

The `location` field is a user-provided, 30-character limited text string for a user to describe his/her current location. However, as this is a free-form text field, users can enter any text desired, even nonsensical or non-existent locations. Some mobile Twitter clients populate this field with the exact geographic coordinates of the user, based on their devices' inbuilt GPS feature.

URL

Different from URL strings in tweet messages, the `url` field lets the user specify a profile URL such as a homepage or a blog that will appear when the user's profile is viewed by others. Originally, personal webpages or personal profile pages on other Web services were used in this field. However, as the user base evolved, the usage of celebrity pages in place of personal profile URLs have since become commonplace [Cormode et al., 2010].

User locale properties

Three pieces of user metadata — containing properties of the user locale — are generated by Twitter for each user account, based on information provided upon account creation, and user activity.

`lang` is the ISO 639-1 language code, specifying the language used by the current user to write messages in. This is mainly used by third-party language-specific clients[4] to indicate the language used in tweets.

`time_zone` (and the corresponding `utc_offset`) stores time zone information for the current user profile; this is specified by the user in the account preferences section. `time_zone` is a string containing the proper name of the time zone, while `utc_offset` represents its offset — in seconds, either positive or negative — from UTC (Coordinated Universal Time).

User behavior summaries

Twitter also lists a set of user behavior statistics, one of which is the number of tweets the user has added to his favorites (`favorites_count`). Also the number of lists the current user is part of (e.g. added to by other Twitter users) is tallied in the `listed_count` metadata item.

The `favorites_count` has been in existence since Twitter's early days; whereas the `listed_count` feature was only present since the introduction of the *Lists* API (discussed earlier in Section 4.1.4).

User activity

Two pieces of metadata are provided by the Twitter API to measure the rate of activity of a user. `statuses_count` stores the total number of tweets the user has composed since the account was created, and `created_at` logs the time of account creation.

User in-degree/out-degree

Twitter's API provides two variables which summarizes a user's social network connections on Twitter. The number of followers the user has (user in-degree in the social graph) and the number of people the user currently *follows* (user out-degree) is provided in the `followers_count` and `friends_count` variables, respectively. It is pertinent to note that the number of mutually followed users (or reciprocal friends as defined by [Kwak et al., 2010] in Section 3.1.2) is not directly available from the API: the two lists of followers and friends have to be manually crawled to find a common subset of users.

Customization

A variety of metadata found in the user domain lets one observe the degree of customization a user has performed on his Twitter user page. These customizations are mainly of use for Twitter (and third-party applications) to customize the display of a user's profile when viewed. This seems trivial at first glance, but the availability of such metadata provides a glimpse into a user's customization behavior throughout his experience with the Twitter service.

[4] As documented in the Twitter API documentation on `search` API results: <`https://dev.twitter.com/docs/api/1/get/search`>

Firstly, the Boolean summary variables `default_profile` and `default_profile_image` states whether the user has changed his profile and profile picture, respectively, from the default settings.

The following are the other metadata variables from the API, describing the customization of a user's profile:

- `profile_image_url`: This is a string containing the URL to the user's profile picture (also known as *online avatar*) that can be seen by other users on Twitter.

- `profile_background_color`: A triplet of two-character hexadecimal values in the format `rrggbb`; specifying the red, green, and blue color components of the profile background respectively.

- `profile_background_image_url`: This is a string containing the URL to the custom background image the user has for his Twitter profile. (As of time of writing, ready-made theme images provided by Twitter contain the substring `/themes/theme`N, where the theme choice is specified by the decimal number N).

- `profile_use_background_image`: A Boolean flag specifying if the user background image is used in place of a flat background color.

- `profile_background_tile`: A Boolean flag specifying if the user background image is tiled.

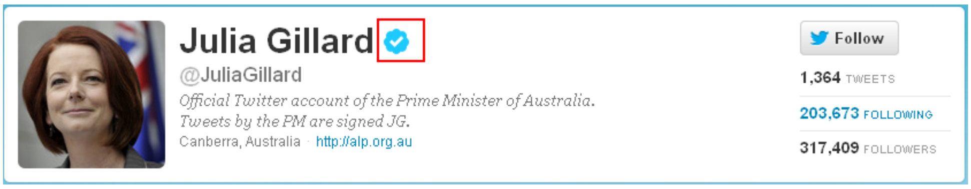
- `profile_text_color`, `profile_link_color`, `profile_sidebar_border_color`, and `profile_sidebar_fill_color`: Triplets of hexadecimal values in the format `rrggbb` (as per the convention of `profile_background_color`) which specifies the color of the profile text, links, sidebar border, and sidebar fill respectively.

User flags

Twitter also stores Boolean flags denoting certain properties of users, which are then presented as user metadata variables accessible through the Twitter API.

One of the important flags available since the launch of Twitter is the `protected` flag: a simple Boolean value which determines if the user account is set to protected. Protected accounts are only viewable to explicitly authorized users but not the general public. As discussed in Chapter 2, throughout this PhD thesis I only use data from accounts which are visible to the general public (i.e. `protected` = `false`).

Near the end of 2009, Twitter has introduced the `verified` flag, which confirms the true identity of a user for a high-profile Twitter account. This means high-profile users, such as celebrities or politicians, can apply to Twitter Inc. to verify their identity as the legitimate owner of an official Twitter account. A usage scenario for this is to allow users to easily identify an official account belonging to say a politician, compared to satirical or parody accounts about himself/herself. Users with the `verified` flag set to `true` will have a special verified icon in his/her Twitter profile page (Figure A.1).



Julia Gillard

@JuliaGillard

Official Twitter account of the Prime Minister of Australia. Tweets by the PM are signed JG.

Canberra, Australia http://alp.org.au

1,364 TWEETS

203,673 FOLLOWING

317,409 FOLLOWERS

Figure A.1: An example of a *verified* Twitter account, in this case, Australian Prime Minister Julia Gillard <`https://twitter.com/JuliaGillard`>. Observe the presence of the Twitter-generated 'verified' icon — the blue shape with a white tick mark within — highlighted with a red square.

Twitter allows the user the option to add a location to all tweets: this preference is reflected in the `geo_enabled` Boolean variable, also introduced the same time as the `verified` flag. This feature allows Twitter to use the location information with GPS-enabled Twitter clients, or location-enabled browsers to tag user's individual messages with the correct geographic location.

Developmental or experimental flags that are also exposed by the Twitter API include:

- `is_translator` flag[5]: Indicates if the user is helping Twitter in its interface translation/localization project (the *Twitter Translation Center*).

- `contributors_enabled` indicates if this account has the *Contributors* experimental feature (which allows multiple users to 'contribute' or post messages as another user).

Observer preferences

Finally, this section details *observer preferences*: settings relating to the target user being queried, in relation to the *observer* (i.e. the currently logged in Twitter user that is consuming the data provided by the API).

As with profile customization, this is used mainly in the design of interfaces or Twitter client applications. As these setting change depending on whichever user account is querying the API, I opine that these settings are only beneficial in research involving user personalization.

- `following`: Boolean value indicating that the observer is *following* the user in question.

- `follow_request_sent`: Boolean value indicating if a 'follow request' has been sent by the observer to the user (if the target user has a protected account).

- `notifications`: Variable specifying notifications from the target user to the observer (deprecated).

- `show_all_inline_media`: Boolean value indicating if the observer prefers to see all media inline — such as linked pictures or video — when viewing the target user using the Twitter web interface.

[5] Clarified in a Twitter Development user group post: <`http://groups.google.com/group/twitter-development-talk/browse_thread/thread/5c10cc4f51d148f2`>

A.2 Sample Raw Metadata Captures

This section contains examples of raw metadata produced by querying the Twitter Streaming API (Section 4.2). This data, in its raw form, is encoded in JavaScript Object Notation for consumption by my experimental Perl script. For readability, this data is then sanitized by removing non-readable encoding, leaving the raw data representation intact.

By listening to the Streaming API, a continuous stream of *messages* is captured, each of them represented as a hashtable data structure (or 'associative memory') containing attributes. It can be seen in the following example that a *message* object consists of a linked *user* object attached to it; the Twitter API automatically enumerates all linked objects (this can also be seen in the linked *entities* object detailing related messages, automatically generated by the API as well).

Table A.1: Sample text dump from a hash table containing the first ten key-value pairs of `source` strings and their corresponding device classes, obtained as a result of Experiment 4.8 .

```
{
    "in_reply_to_status_id_str":null
    "text":"i love him :) haha"
    "in_reply_to_screen_name":null
    "in_reply_to_user_id_str":null
    "id_str":"75097276711841792"
    "contributors":null
    "retweeted":false
    "geo":null
    "truncated":false
    "source":"web"
    "coordinates":null
    "entities":{
        "user_mentions":[]
        "hashtags":[]
        "urls":[]
        }
    "in_reply_to_status_id":null
    "created_at":"Mon May 30 07:12:40 +0000 2011"
    "place":null
    "in_reply_to_user_id":null
    "user":{
        "default_profile_image":false
        "profile_background_image_url":"http:\/\/a2.twimg.com\/profile_background_images\/
            245794547\/156987_1766313596979_1213320019_2070774_4825172_n.jpg"
        "default_profile":false
        "url":null
        "id_str":"37433772"
        "show_all_inline_media":false
        "geo_enabled":false
        "profile_text_color":"521152"
        "follow_request_sent":null
        "profile_sidebar_fill_color":"9c5c7e"
        "followers_count":364
        "profile_image_url":"http:\/\/a0.twimg.com\/profile_images\/1374390097\/
            247308_2132724877032_1213320019_2661721_5319766_n_normal.jpg"
        "description":"apologize if i say everything i dont mean! i'll i care about
        is money and the city that im from -drake"
        "profile_background_tile":true
        "location":"FRM shreveport iACT louisiana"
        "contributors_enabled":false
        "statuses_count":7667
        "screen_name":"renn_djonn"
        "is_translator":false
        "favourites_count":4
        "profile_link_color":"a8054e"
        "listed_count":1
        "lang":"en"
        "verified":false
        "notifications":null
        "created_at":"Sun May 03 15:17:49 +0000 2009"
        "profile_sidebar_border_color":"f5e2e9"
        "protected":false
        "time_zone":"Central Time (US & Canada)"
        "name":"carenn baylor"
        "profile_use_background_image":true
        "friends_count":239
        "id":37433772
        "following":null
        "utc_offset":-21600
        "profile_background_color":"0a090a"
        }
    "retweet_count":0
    "id":75097276711841792
    "favorited":false
}

{
    "in_reply_to_status_id_str":"75096931222831105"
    "text":"@UFC_buddha_MTG lol really ? i didn't notice lol =p"
    "in_reply_to_screen_name":"UFC_buddha_MTG"
    "in_reply_to_user_id_str":"159396508"
    "id_str":"75097276707639296"
    "contributors":null
    "retweeted":false
    "geo":null
    "truncated":false
    "source":"web"
    "coordinates":null
    "entities":{
        "user_mentions":[
            {
                "indices":[0, 15]
                "id_str":"159396508"
                "screen_name":"UFC_buddha_MTG"
                "name":"lloyd dell wagoner"
                "id":159396508
            }
        ]
        "urls":[]
        "hashtags":[]
        }
    "in_reply_to_status_id":75096931222831105
    "created_at":"Mon May 30 07:12:40 +0000 2011"
    "place":null
    "in_reply_to_user_id":159396508
    "user":{
        "default_profile_image":false
        "profile_background_image_url":"http:\/\/a1.twimg.com\/images\/themes\/theme9\/bg.gif"
        "url":null
        "id_str":"165153747"
        "show_all_inline_media":false
        "geo_enabled":false
        "profile_text_color":"666666"
        "follow_request_sent":null
        "profile_sidebar_fill_color":"252429"
        "followers_count":3
        "profile_image_url":"http:\/\/a2.twimg.com\/profile_images\/1365428680\/Photo_00002_normal.jpg"
        "description":""
        "profile_background_tile":false
        "location":""
        "contributors_enabled":false
        "statuses_count":4
        "screen_name":"mintsugerplum"
        "is_translator":false
        "favourites_count":0
        "profile_link_color":"2FC2EF"
        "default_profile":false
        "listed_count":0
        "lang":"en"
        "verified":false
        "notifications":null
        "created_at":"Sat Jul 10 19:18:13 +0000 2010"
        "profile_sidebar_border_color":"181A1E"
        "protected":false
        "time_zone":null
        "name":"Renee Betancourt"
        "profile_use_background_image":true
        "friends_count":1
        "id":165153747
        "following":null
        "utc_offset":null
        "profile_background_color":"1A1B1F"
        }
    "retweet_count":0
    "id":75097276707639296
    "favorited":false
}


{
    "in_reply_to_status_id_str":"75096687370178561"
    "text":"@crystal504sir @babylovez13 @butterzzzzzs @valerierusssian @chrisstanley18
        @angiegoon504 haha I was the 1st one lol"
    "in_reply_to_screen_name":"crystal504sir"
    "in_reply_to_user_id_str":"290233213"
    "id_str":"75097276732805120"
    "contributors":null
    "retweeted":false
    "geo":null
    "truncated":false
    "source":"\u003Ca href=\"http:\/\/twitter.com\/#!\/download\/iphone\" rel=\"nofollow\"
        \u003ETwitter for iPhone\u003C\/a\u003E"
    "coordinates":null
    "entities":{
        "user_mentions":[
            {
                "indices":[0 14]
                "id_str":"290233213"
                "screen_name":"crystal504sir"
                "name":"Crystal Russian"
                "id":290233213
            }
            {
                "indices":[15 27]
                "id_str":"271280025"
                "screen_name":"BabyLovez13"
                "name":"MannaZz"
                "id":271280025
            }
            {
                "indices":[28 41]
                "id_str":"277268110"
                "screen_name":"ButterZZzzZS"
                "name":"they call me gina"
                "id":277268110
            }
            {
                "indices":[42 58]
                "id_str":"237791017"
                "screen_name":"ValerieRusSsian"
                "name":"MissDarkVader"
                "id":237791017
            }
            {
                "indices":[59 74]
                "id_str":"234657721"
                "screen_name":"chrisstanley18"
                "name":"chris stanley "
                "id":234657721
            }
            {
                "indices":[75 88]
                "id_str":"286057527"
                "screen_name":"angiegoon504"
                "name":"angelina goon "
                "id":286057527
            }
            ]
            "urls":[]
            "hashtags":[]
        }
    "in_reply_to_status_id":75096687370178561
    "created_at":"Mon May 30 07:12:40 +0000 2011"
    "place":null
    "in_reply_to_user_id":290233213
    "user":{
        "default_profile_image":false
        "profile_background_image_url":"http:\/\/a0.twimg.com\/images\/themes\/theme1\/bg.png"
        "url":null
        "id_str":"304576240"
        "show_all_inline_media":false
        "geo_enabled":false
        "profile_text_color":"333333"
        "follow_request_sent":null
        "profile_sidebar_fill_color":"DDEEF6"
        "followers_count":10
        "profile_image_url":"http:\/\/a2.twimg.com\/profile_images\/1367523707\/image_normal.jpg"
        "description":"from Today On i dont care what Ppl think about Me so Its up
            to You Follow Me IF You Want Its Up To Youz"
        "profile_background_tile":false
        "location":"Jinx New orleans "
        "contributors_enabled":false
        "statuses_count":281
        "screen_name":"SelenaGizmoz"
        "is_translator":false
        "favourites_count":1
        "profile_link_color":"0084B4"
        "listed_count":0
        "lang":"en"
        "verified":false
        "notifications":null
        "created_at":"Tue May 24 19:05:22 +0000 2011"
        "profile_sidebar_border_color":"C0DEED"
        "protected":false
        "time_zone":null
        "name":"Selenagizmoz"
        "default_profile":true
        "profile_use_background_image":true
        "friends_count":10
        "id":304576240
        "following":null
        "utc_offset":null
        "profile_background_color":"C0DEED"
        }
    "retweet_count":0
    "id":75097276732805120
    "favorited":false
}
```

Appendix B

Listing of `source` Strings and TLDs

This appendix contains data tables from my analysis of raw `source` and `url` strings, discussed in Sections 4.6.3 and 4.7.2.

B.1 Statistics on Device `source` Strings

In Section 4.6.3, recall that the `source` strings from the 7,863,650 tweets in the *10-Gigabyte Dataset* were grouped and ranked by frequency. Using the manual categorization scheme in Experiment 4.8, the top 300 most-frequently-occurring `source` string is annotated with an individual *device class*, defined in Table 4.9. The `source` strings for Twitter clients uniquely targeted towards particular countries (as per Section 5.4) are also identified.

Table B.1 lists the statistics for our aforementioned 300 `source` strings, with some category labels abbreviated for brevity.

Table B.1: Top 300 `source` strings with their annotated device class and frequency ranking.

Rank	`source` string	Category	Region-specific	Frequency	Percentage
1	*web*	Web		2,381,190	30.2810%
2	*Twitter for iPhone*	Mobile		855,886	10.8841%
3	*Twitter for BlackBerry*	Mobile		601,546	7.6497%
4	*Twitter for Android*	Mobile		453,131	5.7623%
5	*UberSocial for BlackBerry*	Mobile		401,827	5.1099%
6	*Mobile Web*	Mobile		314,648	4.0013%
7	*TweetDeck*	Social network int.		266,418	3.3880%
8	*Echofon*	Mobile		189,810	2.4138%
9	*twittbot.net*	Bots		151,623	1.9282%
10	*Keitai Web*	Web	Japan	146,981	1.8691%
11	*twitterfeed*	Feed aggregators		121,378	1.5435%
12	*twicca*	Mobile	Japan	110,564	1.4060%
13	*Plume*	Mobile		104,919	1.3342%
14	*TweetCaster for Android*	Mobile		101,492	1.2906%
15	*txt*	Mobile		87,383	1.1112%
16	*twipple*	Mobile	Japan	69,054	0.8781%
17	*Tumblr*	Web 2.0 int.		64,031	0.8143%
18	*Tweet Button*	Web 2.0 int.		62,435	0.7940%
19	*HootSuite*	Marketing tools		60,891	0.7743%
20	*Facebook*	Web 2.0 int.		49,461	0.6290%

Rank	source string	Category	Region-specific	Frequency	Percentage
21	*UberSocial for Android*	Mobile		41,898	0.5328%
22	*Tween*	Interfaces	Japan	40,459	0.5145%
23	*Twitter for iPad*	Mobile		39,771	0.5058%
24	*www.movatwi.jp*	Mobile	Japan	39,009	0.4961%
25	*Write Longer*	Interfaces		35,636	0.4532%
26	*Twitter for Mac*	Interfaces		30,753	0.3911%
27	*Twipple for iPhone*	Mobile	Japan	29,078	0.3698%
28	*SOICHA*	Social network int.	Japan	28,143	0.3579%
29	*Twidroyd for Android*	Mobile		24,969	0.3175%
30	*Google*	Web 2.0 int.		24,787	0.3152%
31	*dlvr.it*	Feed aggregators		23,937	0.3044%
32	*jigtwi*	Mobile	Japan	23,185	0.2948%
33	*foursquare*	Web 2.0 int.		22,569	0.2870%
34	*Seesmic*	Social network int.		20,503	0.2607%
35	*Twipple for Android*	Mobile	Japan	17,821	0.2266%
36	*Tweetbot for iPhone*	Mobile		17,815	0.2265%
37	*Janetter2*	Interfaces		16,685	0.2122%
38	*TwitBird*	Mobile		14,788	0.1881%
39	*yubitter*	Mobile	Japan	14,773	0.1879%
40	*UberSocial*	Mobile		13,889	0.1766%
41	*m.tweete.net*	Alternate proxies		13,466	0.1712%
42	*Saezuri*	Interfaces	Japan	12,028	0.1530%
43	*EasyBotter*	Bots		11,585	0.1473%
44	*Tweetlogix*	Mobile		11,527	0.1466%
45	*Instagram*	Web 2.0 int.		11,481	0.1460%
46	*Twitcam.com*	Third-party		10,796	0.1373%
47	*YoruFukurou*	Interfaces	Japan	10,030	0.1275%
48	*SocialOomph*	Marketing tools		9,453	0.1202%
49	*HTC Peep*	Mobile		9,048	0.1151%
50	*TweetCaster for iOS*	Mobile		8,914	0.1134%
51	*twtkr*	Interfaces		8,628	0.1097%
52	*Samsung Mobile*	Mobile		8,603	0.1094%
53	*Ovi by Nokia*	Mobile		8,592	0.1093%
54	*Twitpic*	Web 2.0 int.		8,535	0.1085%
55	*Osfoora for iPhone*	Mobile		7,213	0.0917%
56	*witter.softama.com*	Interfaces	Japan	6,809	0.0866%
57	*Tuitwit*	Alternate proxies		6,627	0.0843%
58	*Twittelator*	Mobile		6,454	0.0821%
59	*A.plus for BlackBerry*	Branded		5,995	0.0762%
60	*SocialScope*	Mobile		5,480	0.0697%
61	*Twitterrific*	Interfaces		5,315	0.0676%
62	*LG Phone*	Mobile		5,293	0.0673%
63	*Snaptu*	Third-party		5,225	0.0664%
64	*Gravity!*	Mobile		5,119	0.0651%
65	*GetGlue.com*	Web 2.0 int.		5,041	0.0641%
66	*UberSocial Mobile*	Alternate proxies		4,912	0.0625%
67	*UberSocial for iPhone*	Mobile		4,905	0.0624%
68	*Silver Bird*	Interfaces		4,825	0.0614%
69	*Twittascope*	Third-party		4,769	0.0606%
70	*Dabr*	Alternate proxies		4,450	0.0566%

Rank	source string	Category	Region-specific	Frequency	Percentage
71	*Ustream.TV*	Web 2.0 int.		4,388	0.0558%
72	*Tweet Old Post*	Feed aggregators		4,229	0.0538%
73	*vk.com*	Suspicious		4,176	0.0531%
74	*Tower Heist Takeover*	Games		4,056	0.0516%
75	*Ping.fm*	Social network int.		3,936	0.0501%
76	*Twitter for Windows Phone*	Mobile		3,891	0.0495%
77	*Weather Display*	Web 2.0 int.		3,837	0.0488%
78	*Nimbuzz Mobile*	Mobile		3,734	0.0475%
79	*TweetCaster*	Mobile		3,624	0.0461%
80	*Twiterous*	Alternate proxies		3,583	0.0456%
81	*TwitPal*	Mobile		3,517	0.0447%
82	*Tweet ATOK*	Mobile		3,385	0.0430%
83	*mixi*	Web 2.0 int.	Japan	3,384	0.0430%
84	*Twil2 (Tweet Anytime*	Access gateways	Japan	3,192	0.0406%
85	*nicovideo.jp*	Web 2.0 int.	Japan	3,092	0.0393%
86	*Buffer*	Marketing tools		2,969	0.0378%
87	*Formspring.me*	Web 2.0 int.		2,821	0.0359%
88	*Teewee*	Interfaces		2,794	0.0355%
89	*natetweeting*	Mobile		2,723	0.0346%
90	*twtkr for iPhone*	Mobile		2,717	0.0346%
91	*twitbeam*	Mobile	Japan	2,711	0.0345%
92	*CoTweet*	Marketing tools		2,710	0.0345%
93	*Retwedia.com*	Third-party		2,638	0.0335%
94	*Tweetwawa*	Third-party		2,629	0.0334%
95	*TweetList!*	Mobile		2,618	0.0333%
96	*CitCuit*	Alternate proxies		2,608	0.0332%
97	*Photos on iOS*	Web 2.0 int.		2,606	0.0331%
98	*movatwi.jp*	Mobile	Japan	2,598	0.0330%
99	*TweetList Pro*	Mobile		2,586	0.0329%
100	*TwitIQ*	Interfaces		2,538	0.0323%
101	*Twit for Windows*	Interfaces	Japan	2,531	0.0322%
102	*Twipple Pro for iPhone*	Mobile	Japan	2,509	0.0319%
103	*NetworkedBlogs*	Feed aggregators		2,435	0.0310%
104	*DestroyTwitter*	Interfaces		2,419	0.0308%
105	*LinksAlpha*	Social network int.		2,393	0.0304%
106	*Twuffer*	Marketing tools		2,293	0.0292%
107	*RuTwitPost*	Suspicious		2,289	0.0291%
108	*Camera on iOS*	Web 2.0 int.		2,160	0.0275%
109	*romedes*	Suspicious		2,156	0.0274%
110	*Twitscoop*	Interfaces		2,145	0.0273%
111	*ifttt*	Social network int.		2,100	0.0267%
112	*Panoramic moTweets*	Mobile		2,073	0.0264%
113	*Addictweet*	Mobile		1,946	0.0247%
114	*ALToolbar*	Suspicious		1,903	0.0242%
115	*Favstar.FM*	Third-party		1,899	0.0241%
116	*Paper.li*	Third-party		1,895	0.0241%
117	*WordPress.com*	Feed aggregators		1,887	0.0240%
118	*Tabtter*	Interfaces	Japan	1,811	0.0230%
119	*Twipple for iPad*	Mobile	Japan	1,809	0.0230%
120	*MetroTwit*	Interfaces		1,800	0.0229%

Rank	source string	Category	Region-specific	Frequency	Percentage
121	*twtkr for Android*	Mobile		1,664	0.0212%
122	*TweetCatch.com*	Alternate proxies		1,645	0.0209%
123	*Like My Tweets*	Third-party		1,637	0.0208%
124	*LinkedIn*	Web 2.0 int.		1,615	0.0205%
125	*RockMelt*	Feed aggregators		1,607	0.0204%
126	*UncleUber for Blackberry*	Branded		1,565	0.0199%
127	*The Perfect Quran*	Web 2.0 int.		1,562	0.0199%
128	*yoono*	Social network int.		1,562	0.0199%
129	*Ask.fm*	Social network int.		1,527	0.0194%
130	*MySpace*	Web 2.0 int.		1,524	0.0194%
131	*Tweets60dm*	Mobile		1,524	0.0194%
132	*FC2 Blog Notify*	Feed aggregators		1,477	0.0188%
133	*P3:PeraPeraPrv*	Interfaces	Japan	1,467	0.0187%
134	*twitaddons*	Third-party		1,462	0.0186%
135	*NewsForward for BlackBerry*	Feed aggregators		1,419	0.0180%
136	*Sandaysoft Cumulus*	Feed aggregators		1,418	0.0180%
137	*WordTwit Plugin*	Feed aggregators		1,416	0.0180%
138	*livedoor Blog*	Feed aggregators	Japan	1,410	0.0179%
139	*Social App by dtac*	Social network int.		1,345	0.0171%
140	*Twisuke*	Third-party	Japan	1,338	0.0170%
141	*Movatter*	Mobile	Japan	1,332	0.0169%
142	*twidroyd*	Mobile		1,300	0.0165%
143	*TwitCasting*	Web 2.0 int.		1,237	0.0157%
144	*SKY Androian*	Mobile	Japan	1,215	0.0155%
145	*TwitMania.com*	Alternate proxies		1,197	0.0152%
146	*MOTOBLUR*	Mobile		1,186	0.0151%
147	*TweetMe for iPhone*	Mobile	Japan	1,162	0.0148%
148	*bitly*	Web 2.0 int.		1,134	0.0144%
149	*Yfrog*	Web 2.0 int.		1,129	0.0144%
150	*SocialAdsPro*	Marketing tools		1,126	0.0143%
151	*Azurea for Windows*	Mobile		1,116	0.0142%
152	*rakubo2*	Bots	Japan	1,105	0.0141%
153	*Twibow*	Interfaces	Japan	1,094	0.0139%
154	**(this source string is in a different character set, non-printable in ASCII)**	Mobile		1,093	0.0139%
155	*TwitLonger Beta*	Third-party		1,061	0.0135%
156	*myYearbook Share*	Web 2.0 int.		1,045	0.0133%
157	*Get! PocketVegas*	Games	Japan	994	0.0126%
158	*oi.com*	Feed aggregators		979	0.0124%
159	*TweetMeme*	Third-party		978	0.0124%
160	*Stardoll*	Games		963	0.0122%
161	*WPSyndicator*	Marketing tools		940	0.0120%
162	*vk.com pages*	Suspicious		936	0.0119%
163	*Twittanica*	Alternate proxies		929	0.0118%
164	*Plurk*	Web 2.0 int.		917	0.0117%
165	*Janetter*	Interfaces	Japan	892	0.0113%
166	*YouTube on iOS*	Web 2.0 int.		885	0.0113%
167	*Visibli*	Marketing tools		880	0.0112%
168	*Botize*	Bots		872	0.0111%
169	*TwitBird iPad*	Mobile		862	0.0110%
170	*Tweetie for Mac*	Interfaces		859	0.0109%

Rank	source string	Category	Region-specific	Frequency	Percentage
171	*Seesmic twhirl*	Social network int.		841	0.0107%
172	*TwidroydPRO*	Mobile		838	0.0107%
173	*Sprout Social*	Marketing tools		831	0.0106%
174	*phnx*	Mobile		825	0.0105%
175	*Allvoices.com*	Web 2.0 int.		823	0.0105%
176	*Motorola mTweet*	Mobile		815	0.0104%
177	*Twaimclient*	Alternate proxies		813	0.0103%
178	*Power Twitter*	Interfaces		805	0.0102%
179	*Yahoo!*	Web 2.0 int.		802	0.0102%
180	*Tweetings for iPhone*	Mobile		786	0.0100%
181	*satuspell*	Suspicious		776	0.0099%
182	*weheartit.com*	Web 2.0 int.		776	0.0099%
183	*Hape Esia*	Branded		772	0.0098%
184	*Headliner.fm*	Web 2.0 int.		767	0.0098%
185	*Osfoora HD*	Mobile		762	0.0097%
186	*GO Launcher EX*	Mobile		748	0.0095%
187	*Marci*	Web 2.0 int.		742	0.0094%
188	*Hatena*	Web 2.0 int.	Japan	732	0.0093%
189	*twiroboJP*	Bots	Japan	721	0.0092%
190	*Fwix*	Web 2.0 int.		706	0.0090%
191	*Twipple Pro for Android*	Mobile	Japan	688	0.0087%
192	*Safari on iOS*	Interfaces		685	0.0087%
193	*schmap.it*	Marketing tools		685	0.0087%
194	*Krile2*	Interfaces	Japan	681	0.0087%
195	*KicauMedia*	Alternate proxies		678	0.0086%
196	*The Visitor Widget*	Marketing tools		675	0.0086%
197	*Cidade Maravilhosa: Rio*	Games		661	0.0084%
198	*Twitterrific for Mac*	Interfaces		660	0.0084%
199	*WindowsLive*	Mobile		655	0.0083%
200	*Posterous*	Web 2.0 int.		632	0.0080%
201	*TweetCaster for WP7*	Mobile		628	0.0080%
202	*iTweet.net*	Interfaces		621	0.0079%
203	*fllwrs*	Marketing tools		620	0.0079%
204	*PSP*	Access gateways		619	0.0079%
205	*The Tweeted Times*	Third-party		608	0.0077%
206	*Twit Delay*	Marketing tools		607	0.0077%
207	*Fluppy*	Alternate proxies		605	0.0077%
208	*Tinychat Connector*	Third-party		595	0.0076%
209	*twicli*	Interfaces	Japan	595	0.0076%
210	*Flipboard*	Third-party		594	0.0076%
211	*Futuretweets V2*	Marketing tools		588	0.0075%
212	*Twaitter*	Marketing tools		574	0.0073%
213	*LaterBro.com*	Marketing tools		568	0.0072%
214	*SAM Broadcaster Song Info*	Web 2.0 int.		558	0.0071%
215	*tGadget*	Interfaces		557	0.0071%
216	*Hotpepper*	Web 2.0 int.	Japan	546	0.0069%
217	*LiveJournal.com*	Feed aggregators		544	0.0069%
218	*okinawa_jp_net*	Alternate proxies		543	0.0069%
219	*Kanvaso*	Third-party		541	0.0069%
220	*Securenet Systems Radio Playlist Update*	Marketing tools		534	0.0068%

Rank	source string	Category	Region-specific	Frequency	Percentage
221	*DROID*	Mobile		532	0.0068%
222	*The Sitter for BlackBerry*	Branded		526	0.0067%
223	*hellotxt.com*	Social network int.		522	0.0066%
224	*Sonic Tweet*	Branded		514	0.0065%
225	*SimplyTweet*	Mobile		502	0.0064%
226	*AutoTweet Connector*	Feed aggregators		499	0.0063%
227	*ReverbNation*	Web 2.0 int.		487	0.0062%
228	*shareaholic app*	Web 2.0 int.		482	0.0061%
229	*Triberr*	Marketing tools		480	0.0061%
230	*Daum*	Web 2.0 int.		479	0.0061%
231	*Socially Mobile*	Mobile		473	0.0060%
232	*TweetChat*	Third-party		471	0.0060%
233	*nicovideo.jp live*	Web 2.0 int.	Japan	467	0.0059%
234	*SeesaaBlog*	Feed aggregators	Japan	455	0.0058%
235	*Splitweet*	Marketing tools		454	0.0058%
236	*Twittelator Neue*	Mobile		450	0.0057%
237	*Jobmagic*	Marketing tools		449	0.0057%
238	*www.f-1gp.com/twitters/*	Branded	Japan	449	0.0057%
239	*Imorv*	Suspicious		443	0.0056%
240	*Colotwi*	Games	Japan	439	0.0056%
241	*Tweeker*	Mobile		435	0.0055%
242	*bonheur.rsn.jp/hisaki/android.html*	Mobile	Japan	428	0.0054%
243	*Uber50*	Branded		424	0.0054%
244	*CalTweet*	Marketing tools		423	0.0054%
245	*esavci*	Suspicious		423	0.0054%
246	*Constant Contact*	Marketing tools		420	0.0053%
247	*SNS Analytics*	Marketing tools		417	0.0053%
248	*Dynamic Tweets*	Marketing tools		412	0.0052%
249	*Bullhorn Reach*	Marketing tools		411	0.0052%
250	*bukkake.zz.tc*	Suspicious		406	0.0052%
251	*Xbox*	Access gateways		403	0.0051%
252	*MyAuthAPIProxy*	Alternate proxies		402	0.0051%
253	*25re.com*	Mobile	Japan	397	0.0050%
254	*Miso*	Web 2.0 int.		396	0.0050%
255	*teresti*	Suspicious		393	0.0050%
256	*Yotomo App*	Games		391	0.0050%
257	*Azurea*	Mobile	Japan	390	0.0050%
258	*Princess Punt*	Games		389	0.0049%
259	*appcat.kr/bluebird*	Mobile		388	0.0049%
260	*binReminded*	Third-party		388	0.0049%
261	*TweetCaster for BlackBerry*	Mobile		388	0.0049%
262	*Paradise Island for Android*	Games		384	0.0049%
263	*Raptr*	Web 2.0 int.		384	0.0049%
264	*Twitter MMS*	Mobile		379	0.0048%
265	*twtkr for iPad*	Mobile		379	0.0048%
266	*uplase*	Suspicious		378	0.0048%
267	*Qwitter 5*	Interfaces		376	0.0048%
268	*BlogsFeedNet*	Feed aggregators		370	0.0047%
269	*GroupTweet*	Marketing tools		370	0.0047%
270	*Hotot for Chrome*	Interfaces		370	0.0047%

Rank	`source` **string**	**Category**	**Region-specific**	**Frequency**	**Percentage**
271	*Tweethopper*	Marketing tools		368	0.0047%
272	*SoundCloud*	Web 2.0 int.		367	0.0047%
273	*Newstreamer*	Feed aggregators		364	0.0046%
274	*chirrup.com*	Suspicious		363	0.0046%
275	*BizCaf*	Suspicious		362	0.0046%
276	*Mobile client*	Mobile		362	0.0046%
277	*(olleh talk)*	Mobile		357	0.0045%
278	*embr*	Feed aggregators		357	0.0045%
279	*TwitShepherd*	Mobile	Japan	357	0.0045%
280	*loctouch*	Third-party	Japan	355	0.0045%
281	*s-software.net*	Mobile	Japan	355	0.0045%
282	*Twimbow*	Interfaces		354	0.0045%
283	*Twitcast.me*	Third-party		353	0.0045%
284	*MarketMeSuite*	Marketing tools		349	0.0044%
285	*Realtime love*	Games	Japan	343	0.0044%
286	*Su.pr*	Marketing tools		343	0.0044%
287	*TWEET command*	Bots	Japan	343	0.0044%
288	*Reeder*	Mobile		339	0.0043%
289	*Rock the Vegas for Android*	Games		339	0.0043%
290	*CapturePlay Live*	Branded		338	0.0043%
291	*Digsby*	Social network int.		338	0.0043%
292	*pochitter*	Third-party	Japan	336	0.0043%
293	*Crowy*	Social network int.		336	0.0043%
294	*Pulse News*	Social network int.		334	0.0042%
295	*Loger Simplement*	Web 2.0 int.		332	0.0042%
296	*TheQuotes*	Feed aggregators		330	0.0042%
297	*TweetGrid.com*	Interfaces		329	0.0042%
298	*ra*	Mobile	Japan	328	0.0042%
299	*hiroshi390919_appli*	Feed aggregators	Japan	325	0.0041%
300	*Strictly Tweetbot for Wordpress*	Feed aggregators		323	0.0041%

B.2 Statistics on Top-Level Domains in User url Strings

In Section 4.7.2, I have analyzed the url strings from all 4,491,022 unique users in the *10-Gigabyte Dataset*. Among those users, only 1,536,729 had valid url strings belonging to 338,655 top-level domains (TLDs).

Using the manual categorization scheme in Experiment 4.11, the top 300 most-frequently-occurring TLDs of url strings are annotated with an individual category, defined in Table 4.13. TLDs uniquely targeted towards specific countries (as per Section 4.7.2) are also identified.

Table B.2 lists the statistics for our aforementioned 300 TLDs, with some category labels in short form for brevity.

Table B.2: Top 300 TLDs from user url strings with annotated categories and frequency ranking.

Rank	Website domain	Category	Region-specific?	Occurrences	Percentage of all URLs
1	*facebook.com*	Online social networks		534,042	69.1115%
2	*tumblr.com*	Other microblog		454,899	58.8695%
3	*blogspot.com*	Blog/personal		192,710	24.9390%
4	*twitter.com*	Official Twitter		77,955	10.0883%
5	*youtube.com*	Media sharing		67,395	8.7217%
6	*fc2.com*	Media sharing	Japan	58,507	7.5715%
7	*orkut.com.br*	Online social networks	Brazil	53,315	6.8996%
8	*ameblo.jp*	Blog/personal	Japan	52,609	6.8082%
9	*wordpress.com*	Blog/personal		42,526	5.5034%
10	*pixiv.net*	Media sharing	Japan	36,912	4.7769%
11	*hyves.nl*	Online social networks	Europe	35,920	4.6485%
12	*nicovideo.jp*	Media sharing	Japan	32,038	4.1461%
13	*twilog.org*	Twitter-based media	Japan	29,039	3.7580%
14	*twitpic.com*	Twitter-based media		21,042	2.7231%
15	*formspring.me*	Media sharing		21,019	2.7201%
16	*twpf.jp*	Media sharing	Japan	17,301	2.2390%
17	*ask.fm*	Media sharing		14,717	1.9046%
18	*flavors.me*	Online social networks		14,025	1.8150%
19	*myspace.com*	Online social networks		13,876	1.7957%
20	*heello.com*	Other microblog		13,361	1.7291%
21	*meadiciona.com*	Online social networks	Brazil	13,324	1.7243%
22	*about.me*	Online social networks		12,497	1.6173%
23	*flickr.com*	Media sharing		12,197	1.5784%
24	*bit.ly*	URL shorteners		12,160	1.5737%
25	*hatena.ne.jp*	Media sharing	Japan	10,399	1.3458%
26	*jugem.jp*	Blog/personal	Japan	9,999	1.2940%
27	*vkontakte.ru*	Online social networks	Russia	9,501	1.2295%
28	*livedoor.jp*	Online social networks	Japan	8,988	1.1632%
29	*mixi.jp*	Online social networks	Japan	8,973	1.1612%
30	*cyworld.com*	Online social networks		8,824	1.1419%
31	*soundcloud.com*	Media sharing		8,160	1.0560%
32	*favstar.fm*	Twitter-based media		7,752	1.0032%
33	*reverbnation.com*	Media sharing		7,169	0.9278%
34	*atpages.jp*	Blog/personal	Japan	6,986	0.9041%
35	*google.com*	Online portals		6,685	0.8651%
36	*dclog.jp*	Blog/personal	Japan	6,330	0.8192%
37	*theinterviews.jp*	Online social networks	Japan	6,317	0.8175%
38	*naver.com*	Online portals		6,083	0.7872%
39	*deviantart.com*	Media sharing		5,661	0.7326%
40	*livejournal.com*	Blog/personal		5,622	0.7276%

Rank	**Website domain**	**Category**	**Region-specific?**	**Occurrences**	**Percentage of all URLs**
41	*iddy.jp*	Online social networks	Japan	5,561	0.7197%
42	*nanos.jp*	Media sharing	Japan	5,499	0.7116%
43	*linkedin.com*	Online social networks		5,233	0.6772%
44	*ameba.jp*	Online social networks	Japan	4,988	0.6455%
45	*yahoo.co.jp*	Online portals		4,719	0.6107%
46	*t.co*	URL shorteners		3,819	0.4942%
47	*seesaa.net*	Blog/personal	Japan	3,768	0.4876%
48	*twitterer-wiki.com*	Twitter user indices	Japan	3,662	0.4739%
49	*exblog.jp*	Blog/personal	Japan	3,621	0.4686%
50	*xmbs.jp*	Media sharing	Japan	3,575	0.4626%
51	*shinobi.jp*	Blog/personal	Japan	3,535	0.4575%
52	*yaplog.jp*	Blog/personal	Japan	3,533	0.4572%
53	*webs.com*	Blog/personal		3,491	0.4518%
54	*mblg.tv*	Blog/personal	Japan	3,029	0.3920%
55	*goo.ne.jp*	Informational/news		2,897	0.3749%
56	*exteen.com*	Blog/personal	Thailand	2,800	0.3624%
57	*last.fm*	Media sharing		2,750	0.3559%
58	*sakura.ne.jp*	Media sharing	Japan	2,689	0.3480%
59	*meadd.com*	Blog/personal	Brazil	2,664	0.3448%
60	*wefollow.com*	Twitter user indices		2,632	0.3406%
61	*me2day.net*	Online social networks	South Korea	2,602	0.3367%
62	*fm-p.jp*	Blog/personal	Japan	2,594	0.3357%
63	*skyrock.com*	Online social networks	France	2,551	0.3301%
64	*geocities.jp*	Blog/personal	Japan	2,523	0.3265%
65	*wix.com*	Blog/personal		2,523	0.3265%
66	*tistory.com*	Blog/personal	South Korea	2,440	0.3158%
67	*weheartit.com*	Media sharing		2,424	0.3137%
68	*posterous.com*	Media sharing		2,409	0.3118%
69	*fb.me*	URL shorteners		2,395	0.3099%
70	*weebly.com*	Blog/personal		2,341	0.3030%
71	*tinyurl.com*	URL shorteners		2,248	0.2909%
72	*followfriday.com*	Twitter-based media		2,098	0.2715%
73	*.com*	Top-level Domain		2,094	0.2710%
74	*bandcamp.com*	Media sharing		2,090	0.2705%
75	*cocolog-nifty.com*	Blog/personal	Japan	2,024	0.2619%
76	*pixiv.cc*	Blog/personal	Japan	1,957	0.2533%
77	*blog.me*	Blog/personal	South Korea	1,949	0.2522%
78	*goo.gl*	URL shorteners		1,930	0.2498%
79	*vk.com*	Online social networks	Russia	1,879	0.2432%
80	*crooz.jp*	Blog/personal	Japan	1,822	0.2358%
81	*jimdo.com*	Blog/personal		1,791	0.2318%
82	*so-net.ne.jp*	Online portals	Japan	1,703	0.2204%
83	*etsy.com*	Online sales		1,674	0.2166%
84	*multiply.com*	Online social networks		1,657	0.2144%
85	*twitteris.jp*	Media sharing	Japan	1,648	0.2133%
86	*fotolog.com.br*	Media sharing	Brazil	1,604	0.2076%
87	*koreantweeters.com*	Twitter user indices	South Korea	1,602	0.2073%
88	*fotolog.com*	Media sharing		1,598	0.2068%
89	*gplus.to*	URL shorteners		1,594	0.2063%
90	*twitlonger.com*	Twitter-based media		1,560	0.2019%
91	*atwiki.jp*	Informational/news	Japan	1,559	0.2018%
92	*kaskus.us*	Online portals	Indonesia	1,548	0.2003%
93	*rakuten.co.jp*	Online portals	Japan	1,514	0.1959%
94	*pipa.jp*	Media sharing	Japan	1,507	0.1950%
95	*lastfm.jp*	Media sharing	Japan	1,451	0.1878%
96	*modelmayhem.com*	Online social networks		1,399	0.1810%
97	*egloos.com*	Blog/personal	South Korea	1,389	0.1798%
98	*lyze.jp*	Blog/personal	Japan	1,289	0.1668%
99	*lastfm.com.br*	Media sharing	Brazil	1,282	0.1659%
100	*plurk.com*	Other microblog		1,277	0.1653%

Rank	Website domain	Category	Region-specific?	Occurrences	Percentage of all URLs
101	*fanfiction.net*	Media sharing		1,266	0.1638%
102	*koebu.com*	Media sharing	Japan	1,256	0.1625%
103	*daum.net*	Online portals	South Korea	1,190	0.1540%
104	*wikipedia.org*	Informational/news		1,190	0.1540%
105	*cgiboy.com*	Online social networks	Japan	1,183	0.1531%
106	*peps.jp*	Blog/personal		1,158	0.1499%
107	*youtu.be*	Media sharing		1,117	0.1446%
108	*twpr.jp*	Online social networks	Japan	1,096	0.1418%
109	*booklog.jp*	Media sharing	Japan	1,064	0.1377%
110	*webry.info*	Blog/personal	Japan	1,059	0.1370%
111	*m-pe.tv*	Blog/personal	Japan	1,046	0.1354%
112	*amzn.to*	Online sales		1,041	0.1347%
113	*fblg.jp*	Blog/personal	Japan	991	0.1282%
114	*or.jp*	Top-level Domain	Japan	986	0.1276%
115	*oi.com*	Media sharing		979	0.1267%
116	*ning.com*	Online social networks		971	0.1257%
117	*j.mp*	URL shorteners		963	0.1246%
118	*itsmyurls.com*	URL shorteners		923	0.1194%
119	*luansantana.com.br*	Branding		908	0.1175%
120	*blogg.se*	Blog/personal	Sweden	904	0.1170%
121	*alfoo.org*	Blog/personal	Japan	878	0.1136%
122	*smashblast.co.id*	Branding	Indonesia	873	0.1130%
123	*blog.com*	Blog/personal		871	0.1127%
124	*mypinkfriday.com*	Branding		851	0.1101%
125	*apple.com*	Branding		843	0.1091%
126	*p.tl*	URL shorteners	Japan	814	0.1053%
127	*datpiff.com*	Media sharing		792	0.1025%
128	*yahoo.com*	Online portals		777	0.1006%
129	*decoo.jp*	Blog/personal	Japan	774	0.1002%
130	*meadiciona.com.br*	Online social networks	Brazil	766	0.0991%
131	*xxxxxxxx.jp*	Blog/personal	Japan	751	0.0972%
132	*teacup.com*	Blog/personal	Japan	745	0.0964%
133	*tweetbig.net*	Marketing/bots		745	0.0964%
134	*sblo.jp*	Blog/personal	Japan	737	0.0954%
135	*fwix.com*	Media sharing		705	0.0912%
136	*hotnewhiphop.com*	Media sharing		700	0.0906%
137	*flogao.com.br*	Blog/personal	Brazil	693	0.0897%
138	*nifty.com*	Online portals	Japan	686	0.0888%
139	*paper.li*	Twitter-based media		682	0.0883%
140	*sayat.me*	Online social networks		668	0.0864%
141	*foursquare.com*	Online social networks		661	0.0855%
142	*vimeo.com*	Media sharing		658	0.0852%
143	*bieberfever.com*	Branding		658	0.0852%
144	*.jp*	Top-level Domain	Japan	655	0.0848%
145	*ocn.ne.jp*	Online portals		652	0.0844%
146	*soundclick.com*	Media sharing		650	0.0841%
147	*lastfm.es*	Media sharing	Spain	633	0.0819%
148	*tuenti.com*	Online social networks	Spain	632	0.0818%
149	*ustream.tv*	Media sharing		629	0.0814%
150	*bigcartel.com*	Online portals		625	0.0809%
151	*amazon.co.jp*	Online sales		606	0.0784%
152	*tosp.co.jp*	Media sharing	Japan	568	0.0735%
153	*carbonmade.com*	Online sales		564	0.0730%
154	*migre.me*	URL shorteners	Brazil	559	0.0723%
155	*doorblog.jp*	Blog/personal	Japan	558	0.0722%
156	*fb.com*	Online social networks		557	0.0721%
157	*photozou.jp*	Media sharing	Japan	555	0.0718%
158	*formspring.com*	Media sharing		544	0.0704%
159	*yfrog.com*	Media sharing		541	0.0700%
160	*akahoshitakuya.com*	Blog/personal	Japan	538	0.0696%

Rank	Website domain	Category	Region-specific?	Occurrences	Percentage of all URLs
161	*twittbot.net*	Marketing/bots		533	0.0690%
162	*hotpepper.jp*	Online portals		518	0.0670%
163	*yolasite.com*	Blog/personal		512	0.0663%
164	*tool.ms*	Suspicious		497	0.0643%
165	*pornhub.com*	Adult sites		494	0.0639%
166	*ladygaga.com*	Branding		491	0.0635%
167	*orkut.com*	Online social networks		488	0.0632%
168	*me.com*	Blog/personal		483	0.0625%
169	*looklet.com*	Media sharing		476	0.0616%
170	*dyndns.org*	Blog/personal		474	0.0613%
171	*imdb.com*	Informational/news		470	0.0608%
172	*swasalert.com*	Informational/news		468	0.0606%
173	*land.to*	Blog/personal	Japan	459	0.0594%
174	*lockerz.com*	Online social networks		456	0.0590%
175	*typepad.com*	Blog/personal		455	0.0589%
176	*who-hub.info*	Online social networks		451	0.0584%
177	*carview.co.jp*	Informational/news	Japan	449	0.0581%
178	*twitition.com*	Twitter-based media		444	0.0575%
179	*dothome.co.kr*	Blog/personal	South Korea	438	0.0567%
180	*biglobe.ne.jp*	Online portals	Japan	437	0.0566%
181	*web.id*	Blog/personal		437	0.0566%
182	*livedoor.biz*	Blog/personal	Japan	408	0.0528%
183	*capturEplay.com*	Branding		406	0.0525%
184	*appspot.com*	Media sharing		396	0.0512%
185	*dion.ne.jp*	Blog/personal	Japan	395	0.0511%
186	*nobody.jp*	Blog/personal	Japan	391	0.0506%
187	*schmap.com*	Media sharing		383	0.0496%
188	*blogg.no*	Blog/personal	Norway	381	0.0493%
189	*main.jp*	Blog/personal	Japan	379	0.0490%
190	*tuna.be*	Media sharing		374	0.0484%
191	*blogger.com*	Blog/personal		373	0.0483%
192	*bizcaf.ca*	Online sales		371	0.0480%
193	*redgage.com*	Marketing/bots		369	0.0478%
194	*fc2web.com*	Blog/personal	Japan	366	0.0474%
195	*steamcommunity.com*	Online portals		362	0.0468%
196	*3rin.net*	Blog/personal	Japan	361	0.0467%
197	*skyblog.com*	Blog/personal		361	0.0467%
198	*amazon.com*	Online sales		349	0.0452%
199	*ebay.com*	Online sales		349	0.0452%
200	*twitrax.com*	Media sharing		346	0.0448%
201	*gob.mx*	Informational/news	Mexico	344	0.0445%
202	*google.co.jp*	Online portals		343	0.0444%
203	*wikia.com*	Informational/news		341	0.0441%
204	*weibo.com*	Blog/personal	China	335	0.0434%
205	*mindlessbehavior.com*	Branding		334	0.0432%
206	*loger-simplement.com*	Twitter user indices		333	0.0431%
207	*codysimpson.com*	Branding		331	0.0428%
208	*4shared.com*	Media sharing		329	0.0426%
209	*ivyro.net*	Blog/personal		329	0.0426%
210	*orz.hm*	Blog/personal	Japan	328	0.0424%
211	*go.jp*	Informational/news	Japan	327	0.0423%
212	*furaffinity.net*	Online social networks		321	0.0415%
213	*chu.jp*	Blog/personal	Japan	319	0.0413%
214	*syosetu.com*	Media sharing		318	0.0412%
215	*togetter.com*	Twitter user indices	Japan	318	0.0412%
216	*bbc.co.uk*	Informational/news		317	0.0410%
217	*mbsp.jp*	Blog/personal	Japan	310	0.0401%
218	*purevolume.com*	Media sharing		310	0.0401%
219	*voiceblog.jp*	Media sharing	Japan	308	0.0399%
220	*behance.net*	Media sharing		299	0.0387%

Rank	**Website domain**	**Category**	**Region-specific?**	**Occurrences**	**Percentage of all URLs**
221	*etophot.com*	Blog/personal		297	0.0384%
222	*manutd.com*	Branding		293	0.0379%
223	*twipple.jp*	Branding	Japan	292	0.0378%
224	*.tl.gd*	URL shorteners		292	0.0378%
225	*over-blog.com*	Blog/personal	France	291	0.0377%
226	*xydo.com*	Online social networks		290	0.0375%
227	*blogdetik.com*	Online portals	Indonesia	289	0.0374%
228	*fuckyou.com*	Adult sites		288	0.0373%
229	*live.com*	Blog/personal		287	0.0371%
230	*lastfm.ru*	Media sharing	Russia	287	0.0371%
231	*xrea.com*	Blog/personal	Japan	285	0.0369%
232	*listpipe.com*	Marketing/bots		285	0.0369%
233	*galu-senchu.com*	Online sales		282	0.0365%
234	*hiho.jp*	Blog/personal	Japan	279	0.0361%
235	*stagram.com*	Media sharing		279	0.0361%
236	*arsenal.com*	Branding		278	0.0360%
237	*xrl.us*	URL shorteners		278	0.0360%
238	*dooid.com*	Online social networks		277	0.0358%
239	*globo.com*	Informational/news	Brazil	267	0.0346%
240	*ti-da.net*	Blog/personal	Japan	267	0.0346%
241	*friendfeed.com*	Online social networks		267	0.0346%
242	*ni-moe.com*	Blog/personal	Japan	266	0.0344%
243	*ni-3.net*	Blog/personal		265	0.0343%
244	*or.id*	Top-level Domain		265	0.0343%
245	*iza-yoi.net*	Blog/personal		265	0.0343%
246	*twitvid.com*	Twitter-based media		265	0.0343%
247	*thebomb.com*	Branding		265	0.0343%
248	*jlsofficial.com*	Branding		263	0.0340%
249	*movapic.com*	Media sharing		262	0.0339%
250	*ainsel.org*	Online social networks	Japan	261	0.0338%
251	*ldblog.jp*	Blog/personal	Japan	259	0.0335%
252	*oddfuture.com*	Branding		258	0.0334%
253	*twbirthday.com*	Twitter user indices		256	0.0331%
254	*stickam.jp*	Media sharing	Japan	256	0.0331%
255	*uol.com.br*	Online portals	Brazil	254	0.0329%
256	*cosp.jp*	Online social networks	Japan	253	0.0327%
257	*ratingaddict.com*	Online social networks		253	0.0327%
258	*webnode.com.br*	Blog/personal	Brazil	253	0.0327%
259	*buoyalarm.com*	Informational/news		251	0.0325%
260	*podomatic.com*	Media sharing		249	0.0322%
261	*home.ne.jp*	Online portals		249	0.0322%
262	*socialoomph.com*	Marketing/bots		245	0.0317%
263	*zip.net*	Online portals	Brazil	243	0.0314%
264	*redtube.com*	Adult sites		242	0.0313%
265	*idgaf.com*	Blog/personal		241	0.0312%
266	*cherrybelle.info*	Branding	Indonesia	239	0.0309%
267	*nari-kiri.com*	Blog/personal		238	0.0308%
268	*ph9.jp*	Branding	Japan	237	0.0307%
269	*fotologue.jp*	Media sharing	Japan	235	0.0304%
270	*atm8y.com*	Twitter user indices		234	0.0303%
271	*moo.jp*	Branding	Japan	234	0.0303%
272	*daportfolio.com*	Media sharing		233	0.0302%
273	*ow.ly*	URL shorteners		230	0.0298%
274	*gree.jp*	Online social networks	Japan	229	0.0296%
275	*spillit.me*	Media sharing		229	0.0296%
276	*gnavi.co.jp*	Informational/news	Japan	229	0.0296%
277	*dip.jp*	Blog/personal	Japan	229	0.0296%
278	*jpn.org*	Online portals	Japan	228	0.0295%
279	*realmadrid.com*	Branding		226	0.0292%
280	*wakwak.com*	Online portals	Japan	226	0.0292%

Rank	Website domain	Category	Region-specific?	Occurrences	Percentage of all URLs
281	*client.jp*	Blog/personal	Japan	226	0.0292%
282	*onet.pl*	Online portals	Poland	226	0.0292%
283	*helium.com*	Informational/news		223	0.0289%
284	*selenagomez.com*	Branding		222	0.0287%
285	*blogri.jp*	Blog/personal	Japan	222	0.0287%
286	*blogtalkradio.com*	Media sharing		220	0.0285%
287	*onedirectionmusic.com*	Branding		218	0.0282%
288	*tiny.cc*	URL shorteners		217	0.0281%
289	*go.com*	Online portals		216	0.0280%
290	*polyvore.com*	Media sharing		216	0.0280%
291	*tinami.com*	Media sharing	Japan	216	0.0280%
292	*twitter.com.br*	Official Twitter	Brazil	212	0.0274%
293	*dealsnear.me*	Twitter user indices		212	0.0274%
294	*.is.gd*	URL shorteners		211	0.0273%
295	*efpfanfic.net*	Media sharing	Italy	210	0.0272%
296	*kakuren-bo.com*	Branding	Japan	207	0.0268%
297	*eqla3.com*	Online portals		207	0.0268%
298	*who-hub.org*	Twitter user indices		207	0.0268%
299	*r7.com*	Online portals	Brazil	205	0.0265%
300	*kiwi-us.com*	Blog/personal	Japan	205	0.0265%

Appendix C

Streaming API Socket Programming

This Appendix provides a technical insight into the programming behind the various prototypes covered in Section 5.2.

C.1 Initial *RawStreamer* prototype: Low-level Socket Programming Architecture

For the sake of completeness, the inner working of *RawStreamer* is documented in Algorithm 3.11. *RawStreamer* is an early socket-based prototype to harvest raw data from the Streaming API circa 2009. As the Streaming API was still in its infancy, the availability of dedicated and robust libraries for the Streaming API was limited: this motivated the development of *RawStreamer*.

With the advent of stable library code for accessing Streaming API sockets (discussed in Section 5.2.1), *RawStreamer* was no longer in active use.

I have instead repurposed *RawStreamer* to be a real-time visualization tool for tweets, which allows for pre-processing metadata via the use of inference algorithms (Chapter 4), as they are retrieved from the API. This allows for the display of not only the raw metadata, but any useful statistics such as user demographics and cumulative statistics, for example.

C.2 Refactored Streaming API Low-Level Access

As documented in Section 5.2, I have refactored the Streaming API-based metadata harvester to use the Perl module `AnyEvent::Twitter::Stream` for speed and robustness. The lower-level socket programming code in `AnyEvent::Twitter::Stream`, responsible for handling raw data transmission via sockets, is represented in code as a `streamer` object that is initialized at run-time.

The low-level mechanism to read data from an open socket to the Streaming API — similar to those used in `AnyEvent::Twitter::Stream` (i.e. the *streamobject* instance in Algorithm 5.7) — is described in Algorithm 3.12.

Algorithm 3.11 Pseudocode detailing the logic for *RawStreamer*, which has since been deprecated. It is currently used as a legacy ‘toy’ application for Streaming API visualization.

```
initialize $SDLfont$ text object for visualization
initialize $SDLpicture$ object for visualization
$socket$ ← TCP socket connection to stream.twitter.com
if $socket$ connection failed then
    return
end if
disable buffering for $socket$                              ▷ to ensure data continuity
wait for raw data from $socket$
for all $byte$ retrieved from socket do                              ▷ begin event loop
    $buffer$ ← $buffer$ +$byte$
    $SDLevent$ ← handle SDL input events
    if $SDLevent$ indicates exit keypress then
        return
    else if $SDLevent$ indicates pause keypress then
        toggle $pause$ flag
    end if
    if $buffer$ is not a complete tweet record (a full JSON object) yet then
        next                                           ▷ next iteration of event loop
    end if
    sanitize non-ASCII characters in $buffer$
    $jsonobject$ ← JSON-decode $buffer$
    for all user-selected inferences and metadata of interest do
        apply inference algorithm (e.g. gender) to a given field in $jsonobject$
        append the given algorithm's output to $summarystring$
    end for
    append $scrollingtext$ buffer with current tweet info in $jsonobject$
    $refreshcycle$ ← $refreshcycle + 1$
                          ▷ redraw display only during refresh cycle to avoid flicker/lag
    if $pause$ = false and $refreshcycle$ = update interval  then
        generate $summarystring$ with $SDLfont$
        generate $scrollingtext$ with $SDLfont$
        update $SDLdisplay$
        reset $refreshcycle$
    end if
    append record to $filebuffer$
          ▷ append current record to buffer; dump data to disk only after buffer is full
    if $filebuffer$ is full then
        write $filebuffer$ to file on disk
        flush $filebuffer$
    end if
end for
```

Algorithm 3.12 Low-level mechanism for establishing a Streaming API socket connection and consuming raw data in real-time.

```
socket ← create SSL socket connection to Twitter API at https://stream.twitter.com
verify socket connection
send socket a request for the sample API method
send socket credentials: username and password
read connection header from socket
if connection failed then
    return
end if
wait for raw data
for all byte retrieved from socket do
    buffer ← buffer + byte
    if buffer contains a whole tweet record (a full JSON object) then
        record ← extract metadata records by JSON-decoding buffer
        signal to ‘consumer’ program that a full record is ready
        pass record object to be used by ‘consumer’ program
        clear buffer
    end if
end for
```

References

Abbas, O. A. [2008]. Comparisons between data clustering algorithms, *The International Arab Journal of Information Technology* **5**(3): 320–325.

ABI Research [2012]. 3% of Users Account for One-Fifth of All Money Spent on Mobile Apps, Available at: <`http://www.abiresearch.com/press/3-of-users-account-for-one-fifth-of-all-money-spen`>.

Abraham, L. B., Mörn, M. P. and Vollman, A. [2010]. Women on the web: How women are shaping the internet, *Technical report*, comScore, Inc. Available at: <`http://www.comscore.com/Insights/Presentations_and_Whitepapers/2010/Women_on_the_Web_How_Women_are_Shaping_the_Internet`>.

Abrol, S. and Khan, L. [2010]. TWinner: Understanding news queries with geo-content using Twitter, *Proc. GIR'10*.

Adams, W. L. [2011]. Were Twitter or BlackBerrys used to fan flames of London's riots?, *Time Magazine. Available at: <`http://www.time.com/time/world/article/0,8599,2087337,00.html`>* **Aug 28**.

Adler, S. [1999]. The Slashdot Effect: An Analysis of Three Internet Publications, Available at: <`http://ssadler.phy.bnl.gov/adler/SDE/SlashDotEffect.htm`>.

Alhoniemi, E., Himberg, J., Parhankangas, J. and Vesanto, J. [2005]. SOM Toolbox, Available at: <`http://www.cis.hut.fi/somtoolbox/`>.

Anderberg, M. [1973]. *Cluster analysis for applications*, Probability and mathematical statistics, Academic Press.

André, P., Schraefel, M., Dix, A., White, R., Bernstein, M. and Luther, K. [2010]. Designing for schadenfreude (or, how to express well-being and see if you're boring people), *Proc. CHI 2010 Workshop on Microblogging.*

Arbesman, S. [2004]. The Memespread Project: An initial analysis of the contagious nature of information in social networks., Available at: <`http://www.arbesman.net/memespread.pdf`>.

Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. [2007]. Mining the Blogosphere: Age, gender and the varieties of selfexpression, *First Monday* **12**(9).

Australian Bureau of Statistics [2009]. 3101.0 - Australian Demographic Statistics, Dec 2008. National Statistics, Available at: <`http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0/`>.

Australian Energy Market Operator [2012]. AEMO average price tables, Available at: <`http://www.aemo.com.au/electricity/NEM-Data/Average-Price-Tables.aspx`>.

Bação, F., Lobo, V. and Painho, M. [2005]. Self-organizing maps as substitutes for k-means clustering, *Proc. ICCS'05*, ICCS'05, Springer-Verlag, pp. 476–483.

Bailey, C., Smith, B. and Burkert, B. [2009]. *#hashtags* - what's happening right now on Twitter, Available at: <`http://hashtags.org`>.

Banerjee, N., Chakraborty, D., Dasgupta, K., Mittal, S., Joshi, A., Nagar, S., Rai, A. and Madan, S. [2009]. User interests in social media sites: an exploration with micro-blogs, *Proc. CIKM '09*, pp. 1823–1826.

Barabási, A.-L. and Albert, R. [1999]. Emergence of scaling in random networks, *Science* **286**(5439): 509–512.

Barracuda Networks, Inc. [2010]. Barracuda Labs 2010 midyear security report, *Technical report*, Barracuda Networks, Inc. Available at: <`http://www.barracudalabs.com/downloads/BarracudaLabs2010MidyearSecurityReport.pdf`>.

Battestini, A., Setlur, V. and Sohn, T. [2010]. A large scale study of text messaging use, *Proc. MobileHCI10*, pp. 229–238.

Baumer, E. and Leis, A. [2010]. Minimalists and zealots: Genres of participation in following on Twitter, *Proc. CHI 2010 Workshop on Microblogging.*

Beaumont, C. [2008]. Mumbai attacks: Twitter and Flickr used to break news, The Daily Telegraph.

Bentley, F. and Metcalf, C. [2009]. The use of mobile social presence, *IEEE Pervasive Computing* **8**(4): 35–41.

Bernstein, M., Kairam, S., Suh, B., Hong, L. and Chi, E. H. [2010]. A torrent of tweets: Managing information overload in online social streams, *Proc. CHI 2010 Workshop on Microblogging.*

Beutler, L. E., Reyes, G., Franco, Z. and Housley, J. [2006]. The need for proficient mental health professionals in the study of terrorism, *in* B. Bongar (ed.), *Psychology Of Terrorism*, Oxford University Press, Melbourne, Australia, pp. 32–52.

Bezdek, J. and Pal, N. [1998]. Some new indexes of cluster validity, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **28**(3): 301–315.

Blei, D. M., Ng, A. Y. and Jordan, M. I. [2003]. Latent Dirichlet Allocation, *J. Mach. Learn. Res* **3**: 993–1022.

Bloch, M. and Carter, S. [2009]. Twitter chatter during the Super Bowl, The New York Times. Available at: <`http://www.nytimes.com/interactive/2009/02/02/sports/20090202_superbowl_twitter.html`>.

Blossom, J. [2009]. *Content Nation: Surviving and Thriving as Social Media Changes Our Work, Our Lives, and Our Future*, Wiley.

Böhringer, M. and Gluchowski, P. [2009]. Aktuelle Schlagwörter: Microblogging (translated), *Informatik Spektrum* **32**(6): 506–510.

Bollen, J., Pepe, A. and Mao, H. [2009]. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, *Technical report*, School of Informatics and Computing, Indiana University - Bloomington.

Boyd, D. and Ellison, N. B. [2007]. Social network sites: Definition, history, and scholarship, *Journal of Computer-Mediated Communication* **13**: 210–230.

Boyd, D., Golder, S. and Lotan, G. [2010]. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter, *Proc. HICCS-43*.

Bozkir, A. S., Mazman, S. G. and Sezer, E. A. [2010]. Identification of user patterns in social networks by data mining techniques: Facebook case, *Proc. IMCW 2010*, p. 145153.

Brynjolfsson, E., Hu, Y. J. and Smith, M. D. [2006]. From Niches to Riches: Anatomy of the Long Tail, *MIT Sloan Management Review* **47**(4): 67–71.

Burns, A. and Eltham, B. [2009]. Twitter Free Iran: An evaluation of Twitter's role in public diplomacy and information operations in Iran's 2009 election crisis, *Proc. Communications Policy & Research Forum 2009*, University of Technology, Sydney.

Cashmore, P. [2009a]. Mashable: Jakarta bombings — Twitter user first on the scene, Available at: <`http://mashable.com/2009/07/16/jakarta-bombings-twitter/`>.

Cashmore, P. [2009b]. Mashable: TwitterHIT: Turning Twitter into a junk traffic exchange, Available at: <`http://mashable.com/2009/05/16/twitterhit/`>.

Chen, G., Banerjee, N., Jaradat, S., Tanaka, T., Ko, M. and Zhang, M. [2002]. Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data, *Statistica Sinica* **12**: 241–262.

Cheong, M. [2009]. 'What are you Tweeting about?': A survey of Trending Topics within the Twitter community, *Technical Report 2009/251*, Clayton School of Information Technology, Monash University.

Cheong, M. [2010]. Commonalities in emergent behavior of Twitter messages across topics. (Unpublished).

Cheong, M. and Lee, V. [2009]. Integrating web-based intelligence retrieval and decision-making from the Twitter Trends knowledge base, *Proc. CIKM 2009 Co-Located Workshops: SWSM 2009*, pp. 1–8.

Cheong, M. and Lee, V. [2010a]. Dissecting Twitter: A review on current microblogging research and lessons from related fields, *From Sociology to Computing in Social Networks: Theory, Foundations and Applications*, Vol. 1 of *Lecture Notes in Social Networks*, Springer-Verlag, pp. 343 – 362.

Cheong, M. and Lee, V. [2010b]. A study on detecting patterns in Twitter intra-topic user and message clustering, *Proc. ICPR 2010*, pp. 3125–3128.

Cheong, M. and Lee, V. [2010c]. *Twitmographics*: Learning the emergent properties of the Twitter community, *From Sociology to Computing in Social Networks: Theory, Foundations and Applications*, Vol. 1 of *Lecture Notes in Social Networks*, Springer-Verlag, pp. 323–342.

Cheong, M. and Lee, V. [2010d]. ”Twittering for Earth”: A study on the impact of microblogging activism on Earth Hour 2009 in Australia, *Proc. ACIIDS 2010*.

Cheong, M. and Lee, V. [2011]. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter, *Information Systems Frontiers* **13**(1): 45–59.

Cheong, M. and Ray, S. [2011]. A literature review of recent microblogging developments, *Technical Report 2011/263*, Clayton School of Information Technology, Monash University.

Cheong, M., Ray, S. and Green, D. [2012a]. Interpreting the 2011 London Riots from Twitter metadata, *Proc. SoCPaR 2012*.

Cheong, M., Ray, S. and Green, D. [2012b]. Large-scale socio-demographic pattern discovery on microblog metadata, *Proc. SoCPaR 2012*.

Choudhury, M. D., Sundaram, H., John, A. and Seligmann, D. D. [2008]. Can blog communication dynamics be correlated with stock market activity?, *Proc. HT’08*.

Clark, J. [2009]. 9/11 pager data visualization, Available at: <`http://neoformix.com/2009/Sep11PagerData.html`>.

Claster, W., Dinh, H. and Cooper, M. [2010]. Naïve Bayes and unsupervised artificial neural nets for Cancun tourism social media data analysis, *Proc. NaBIC 2010*.

Collins, T. [2009]. *The Little Book of Twitter*, Michael O’Mara Books Ltd.

Comm, J. [2009]. *Twitter power: how to dominate your market one tweet at a time*, Wiley, Hoboken, NJ.

Cormode, G., Krishnamurthy, B. and Willinger, W. [2010]. A manifesto for modeling and measurement in social media, *First Monday* **15**(9). Available at: <`http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3072/2601`>.

Couronné, T., Beuscart, C. and Chamayou, C. [2009]. Self-organizing map and social networks: Unfolding online social popularity., *Proc. 24th International Symposium on Computer and Information Sciences.*

Cridland, J. [2011]. Mapping the Riots, Available at: <`http://james.cridland.net/blog/mapping-the-riots/`>.

Daly, E. and Orwant, J. [2003]. Text::GenderFromName - search.cpan.org, Available at: <`http://search.cpan.org/~edaly/Text-GenderFromName-0.32/GenderFromName.pm`>.

Davies, D. L. and Bouldin, D. W. [1979]. A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-1**(2): 224–227.

Dawkins, R. [1989]. *The Selfish Gene*, 2 edn, Oxford University Press.

Dearman, D., Kellar, M. and Truong, K. N. [2008]. An examination of daily information needs and sharing opportunities, *Proc. CSCW 2008*, pp. 679–688.

Deboeck, G. and Kohonen, T. (eds) [1998]. *Visual Explorations in Finance: with Self-Organizing Maps*, Springer-Verlag New York, Inc.

Donath, J., Dragulescu, A., Zinman, A., Viégas, F. and Xiong, R. [2010]. Data portraits, *Proc. SIGGRAPH '10*, pp. 375–383.

Du, H., Carroll, J. and Rosson, M. B. [2010]. Public micro-blogging in classrooms: Towards an active learning environment, *Proc. CHI 2010 Workshop on Microblogging.*

Dunlap, J. C. and Lowenthal, P. R. [2009]. Tweeting the night away: Using Twitter to enhance social presence, *Journal of Information Systems Education* **20**(2).

Ebner, M., Lienhardt, C., Rohs, M. and Meyer, I. [2010]. Microblogs in higher education - a chance to facilitate informal and process-oriented learning?, *Computers & Education* **55**(1): 92 – 100.

Ebner, M. and Schiefner, M. [2008]. Microblogging - more than fun?, *Proc. IADIS Mobile Learning Conference 2008*, pp. 155–159.

Ehrlich, K. and Shami, N. [2010]. Microblogging inside and outside the workplace, *Proc. ICWSM 2010.*

Ems, L. [2010]. Twitter use in Iranian, Moldovan and G-20 Summit protests presents new challenges for governments, *Proc. CHI 2010 Workshop on Microblogging.*

Entous, A. [2009]. U.S. military reviews use of Twitter, other sites, Reuters Inc. Available at: <`http://www.reuters.com/article/technologyNews/idUSTRE5735C720090804`>.

Erickson, I. [2008]. The translucence of Twitter, *Proc. Ethnographic Praxis in Industry Conference 2008*, pp. 58–72.

Eudaptics Software GmbH [2005]. *Viscovery Profiler Version 5.0*, Eudaptics Software GmbH, Kupelwiesergasse 27, 1130 Vienna, Austria, Europe.

Finkel, R. A. and Bentley, J. L. [1974]. Quad Trees: a data structure for retrieval on composite keys, *Acta Informatica* **4**: 1–9.

Fleishman, J. [2009]. Mideast hanging on every text and tweet from Iran, Los Angeles Times. Available at: <`http://articles.latimes.com/2009/jun/17/world/fg-iran-image17`>.

Flor, N. V. [2000]. Web business engineering: Memetic marketing, Available at: <`http://www.informit.com/articles/article.aspx?p=19996`>.

Fry, J. [2008]. A web page of one's own, The Wall Street Journal. Available at: <`http://online.wsj.com/article/SB121562102257039585.html`>.

Fukuhara, T., Murayama, T. and Nishida, T. [2005]. Analyzing concerns of people using weblog articles and real world temporal data, *Proc. WWW 2005.*

Gallagher, A. and Chen, T. [2008]. Estimating age, gender, and identity using first name priors, *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pp. 1–8.

Gladwell, M. [2002]. *The Tipping Point: How Little Things Can Make a Big Difference*, Back Bay Books, New York, NY.

Goldstein, D. G. and Goldstein, D. C. [2006]. Profiting from the long tail, *Harvard Business Review* **84**(6): 24–28.

Golovchinsky, G. and Efron, M. [2010]. Making sense of Twitter Search, *Proc. CHI 2010 Workshop on Microblogging.*

Goolsby, R. [2009]. Lifting elephants: Twitter and blogging in global perspective, *in* H. Liu (ed.), *Social Computing and Behavioral Modeling*, Springer-Verlag, pp. 1–7.

Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. [2005]. The predictive power of online chatter, *Proc. SIGKDD 2005.*

Gruhl, D., Liben-Nowell, D., Guha, R. and Tomkins, A. [2004]. Information diffusion through blogspace, *Proc. WWW 2004*, pp. 491–501.

Gupta, N. and Dey, L. [2010]. Detection and characterization of anomalous entities in social communication networks, *Proc. ICPR'10*, pp. 738–741.

Guy, M., Earle, P., Ostrum, C., Gruchalla, K. and Horvath, S. [2010]. Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies, *Advances in Intelligent Data Analysis IX*, Springer-Verlag, pp. 42–53.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. [2001]. On clustering validation techniques, *Intelligent Information Systems Journal* **17**: 107–145.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. [2009]. The WEKA data mining software: An update, *SIGKDD Explorations* **11**(1).

Halliday, J. [2012]. UK riots 'made worse' by rolling news, BBM, Twitter and Facebook, *The Guardian* **March 28**.

Halvey, M. J. and Keane, M. T. [2007]. An assessment of tag presentation techniques, *Proc. WWW '07*, pp. 1313–1314.

Harris, M. [2008]. Barack to the future, *Engineering & Technology* **3**(20): 25.

Hartigan, J. [1975]. *Clustering algorithms*, Wiley series in probability and mathematical statistics: Applied probability and statistics, Wiley.

Hazlewood, W., Makice, K. and Ryan, W. [2008]. Twitterspace: A co-developed display using Twitter to enhance community awareness, *Proc. PDC '08*.

Heil, B. and Piskorski, M. [2009]. New twitter research: Men follow men and nobody tweets, Available at: <`http://blogs.harvardbusiness.org/cs/2009/06/new_twitter_research_men_follo.html`>.

Herring, S., Scheidt, L., Bonus, S. and Wright, E. [2004]. Bridging the gap: A genre analysis of weblogs, *Proc. 37th Hawaii International Conference on System Sciences*, pp. 1–11.

Hodge, K. [2000]. It's all in the memes, Available at: <`http://www.guardian.co.uk/science/2000/aug/10/technology`>.

Honeycutt, C. and Herring, S. [2009]. Beyond microblogging: Conversation and collaboration via Twitter, *Proc. 42nd Hawaii International Conference on System Sciences*, pp. 1–10.

Horn, C. [2010]. *Analysis and classification of Twitter messages*, Master's thesis, Graz University of Technology.

Huang, J., Thornton, K. M. and Efthimiadis, E. N. [2010]. Conversational tagging in Twitter, *Proc. HT10*.

Huberman, B., Romero, D. and Wu, F. [2008a]. Social networks that matter: Twitter under the microscope, *Technical report*, Social Computing Laboratory, HP Labs. Available at: <`http://ssrn.com/abstract=1313405`>.

Huberman, B., Romero, D. and Wu, F. [2008b]. Social networks that matter: Twitter under the microscope, *First Monday* **14**(1).

Hughes, A. and Palen, L. [2009]. Twitter adoption and use in mass convergence and emergency events, *Proc. 6th International ISCRAM Conference*.

Humphreys, L. [2010]. Historicizing microblogging, *Proc. CHI 2010 Workshop on Microblogging.*

Humphreys, L., Gill, P. and Krishnamurthy, B. [2010]. How much is too much? Privacy issues on Twitter, *Proc. ICA 2012.*

Jain, A. and Dubes, R. [1988]. *Algorithms for Clustering Data*, Prentice Hall advanced reference series, Prentice Hall.

Jamali, M. and Abolhassani, H. [2007]. Using self organizing map to infer communities in weblogs' social network, *Proc. European Conference on Data Mining (IADIS 2007).*

Jansen, B. J., Zhang, M., Sobel, K. and Chowdury, A. [2009a]. Micro-blogging as online word of mouth branding, *Proc. CHI 2009.*

Jansen, B. J., Zhang, M., Sobel, K. and Chowdury, A. [2009b]. Twitter power: Tweets as electronic word of mouth, *Journal of ASIS&T* **60**(9): 1–20.

Java, A., Song, X., Finin, T. and Tsen, B. [2009]. Why we Twitter: An analysis of a microblogging community, *Proc. 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, Springer-Verlag, pp. 118–138.

Jennings, R., Nahum, E., Olshefski, D., Saha, D., Shae, Z.-Y. and Waters, C. [2006]. A study of Internet instant messaging and chat protocols, *IEEE Network* **20**(4): 16–21.

Joinson, A. [2008]. Looking at, looking up or keeping up with people?: Motives and use of Facebook, *Proc. CHI 2008*, pp. 1027–1036.

Jones, R., Kumar, R., Pang, B. and Tomkins, A. [2007]. "I Know What You Did Last Summer" — query logs and user privacy, *Proc. CIKM 2007*, pp. 909–914.

Jungherr, A. [2009]. The DigiActive guide to Twitter for activism. Available at: <`http://www.digiactive.org/wp-content/uploads/digiactive_twitter_guide_v1-0.pdf`>.

Jungherr, A. [2010]. Twitter in politics: Lessons learned during the German Superwahljahr 2009, *Proc. CHI 2010 Workshop on Microblogging.*

Kaufman, S. J. and Chen, J. [2010]. Where we Twitter, *Proc. CHI 2010 Workshop on Microblogging.*

Kim, D., Jo, Y., Moon, I.-C. and Oh, A. [2010]. Analysis of Twitter Lists as a potential source for discovering latent characteristics of users, *Proc. CHI 2010 Workshop on Microblogging.*

Kim, W., Jeong, O.-R. and Lee, S.-W. [2010]. On social websites, *Information Systems* **35**: 215236.

Kireyev, K., Palen, L. and Anderson, K. [2009]. Applications of topics models to analysis of disaster-related Twitter data, *Proc. NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*.

Kohonen, T. [1988]. *Self-Organization and Associative Memory*, Springer, Berlin.

Krikorian, R. [2010]. Map of a Twitter status object, Available at: <`http://datasift.com/a/wp-content/themes/datasift/images/tweet_diagram.pdf`>.

Krishnamurthy, B. [2009]. A measure of Online Social Networks (Invited paper), *Proc. COMSNETS'09*.

Krishnamurthy, B., Gill, P. and Arlitt, M. [2008]. A few chirps about Twitter, *Proc. WOSN'08*, pp. 19–24.

Kumar, R., Mahdian, M. and McGlohon, M. [2010]. Dynamics of conversations, *Proc. KDD10*.

Kwak, H., Lee, C., Park, H. and Moon, S. [2010]. What is Twitter, a social network or a news media?, *Proc. WWW 2010*.

Lawler, J. P. and Molluzzo, J. C. [2011]. A survey of first-year college student perceptions of privacy in social networking, *Journal of Computing Sciences in Colleges* **26**(3).

Lee, K., Caverlee, J. and Webb, S. [2010]. Uncovering social spammers: social honeypots + machine learning, *Proc. SIGIR'10*, pp. 435–442.

Levinson, P. [2009]. *New New Media*, Allyn & Bacon.

Li, H. [2005]. *Data Visualization Of Asymmetric Data Using Sammon Mapping and Applications of Self-Organizing Maps*, PhD thesis, Faculty of the Graduate School, The University of Maryland.

Lin, Y.-R., Tolentino, L. and Kelliher, A. [2010]. Tweeting globally, acting locally: Booming and sustaining disability awareness through twitter, *Proc. CHI 2010 Workshop on Microblogging*.

Ling, R. [2005]. The sociolinguistics of SMS: An analysis of SMS use by a random sample of Norwegians, *Computer Supported Cooperative Work*, Vol. 31 of *Mobile Communications*, Springer, London, pp. 335–349.

Lloyd, S. P. [1982]. Least squares quantization in PCM, *IEEE Trans. Inf. Theory* **28**(2): 129–136.

Lock, A. [2011]. Insurers say London riot losses "well over £100m", *City A.M.* **August 9**.

Lollicode SARL [2009]. Twitscoop - stay on top of Twitter!, Available at: <`http://www.twitscoop.com/`>.

Longueville, B. D., Smith, R. S. and Luraschi, G. [2009]. "OMG, from here, I can see the flames!": A use case of mining Location Based Social Networks to acquire spatiotemporal data on forest fires, *Proc. LBSN '09*, pp. 73–80.

Luhn, H. P. [1958]. The automatic creation of literature abstracts, *IBM Journal of Research Development* **2**(2): 159–165.

MacQueen, J. B. [1967]. Some methods for classification and analysis of multivariate observations, *in* L. M. L. Cam and J. Neyman (eds), *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, pp. 281–297.

Mageswari, M. and Goh, L. [2009]. Nizar is Perak MB: Details of the court ruling, *The Star* **May 11**.

Makice, K. [2009a]. Phatics and the design of community, *Proc. CHI 2009*, pp. 3133–3136.

Makice, K. [2009b]. *Twitter API: Up and Running*, O'reilly Media.

Mankoff, J., Matthews, D., Fussell, S. and Johnson, M. [2007]. Leveraging social networks to motivate individuals to reduce their ecological footprints, *Proc. HICSS-40*, pp. 1–10.

Marsden, G. [2002]. Using HCI to leverage communication technology, *Interactions of the ACM* **10**(2): 48–55.

Martin, R. [2009]. CNET Asia Blogs: Tokyo Shift — WWDC and the iPhone 3GS, Available at: <`http://asia.cnet.com/blogs/tokyo-shift/post.htm?id=63011359`>.

Mathioudakis, M. and Koudas, N. [2010]. TwitterMonitor: Trend detection over the Twitter stream, *Proc. SIGMOD10*.

May, J. [2011]. Burning issues, *The Age* **August 13**: 15–17.

Mayer, M. [2009]. *What The Trend?*, Available at: <`http://www.whatthetrend.com`>.

McAllister, B. [2010]. Why "the conversation" isn't necessarily a conversation, *Interactions* **17**(5): 19–21.

McFedries, P. [2009]. *Twitter: tips, tricks, and tweets*, Wiley, Indianapolis, IN.

McNely, B. J. [2009]. Backchannel persistence and collaborative meaning-making, *Proc. SIGDOC'09*.

Meikle, J. and Jones, S. [2011]. UK riots: More than 1,000 arrests strain legal system to the limit, *The Guardian* **August 10**.

Mendoza, M., Poblete, B. and Castillo, C. [2010]. Twitter under crisis: Can we trust what we RT?, *Proc. SOMA '10*.

Merelo-Guervs, J. J., Prieto, B., Prieto, A., Romero, G., Valdivieso, P. C. and Tricas, F. [2004]. Clustering web-based communities using self-organizing maps, *Proc. IADIS International Conference Web Based Communities 2004*.

Metaxas, P. T. and Mustafaraj, E. [2010]. From obscurity to prominence in minutes: Political speech and realtime search, *Proc. WebSci10*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. [1990]. WordNet: An on-line lexical database, *International Journal of Lexicography* **3**: 235–244.

Milligan, G. W. [1996]. Clustering and classification, *in* P. Arabie, L. Hubert and G. De Soete (eds), *Clustering Validation: Results and Implications for Applied Analyses*, World Scientific, pp. 341–375.

Mischaud, E. [2007]. *Twitter: Expressions of the whole self*, Master's thesis, London School of Economics and Political Science.

Mitra, S. and Acharya, T. [2003]. *Data Mining: Multimedia, Soft Computing, and Bioinformatics*, John Wiley & Sons, Inc.

Moh, T.-S. and Murmann, A. J. [2010]. Can You Judge a Man by His Friends? - Enhancing spammer detection on the Twitter microblogging platform using friends and followers, *Information Systems, Technology and Management*, Vol. 54 of *Communications in Computer and Information Science*, Springer Berlin Heidelberg, pp. 210–220.

Moore, R. J. [2009]. Twitter data analysis: An investors perspective, Available at: <`http://www.techcrunch.com/2009/10/05/twitter-data-analysis-an-investors-perspective/`>.

Motoyama, M., McCoy, D., Levchenko, K. and Voelker, G. M. [2011]. Dirty jobs: The role of freelance labor in web service abuse, *Proc. USENIX 2011*.

Mungiu-Pippidi, A. and Munteanu, I. [2009]. Moldova's "Twitter Revolution", *Journal of Democracy* **20**(3): 136–142.

Musil, S. [2008]. U.S. Army warns of twittering terrorists, CBS Interactive Inc. Available at: <`http://news.cnet.com/8301-1009_3-10075487-83.html`>.

Naaman, M., Boase, J. and Lai, C. [2010]. Is it Really About Me? Message content in social awareness streams, *Proc. CSCW 2010*.

National Electricity Market Management Company Limited [2009]. NEMMCO, Available at: <`http://www.nemmco.com.au/`>.

Natural Earth [2012]. Available at: <`http://www.naturalearthdata.com`>.

Neria, Y., Suh, E. and Marshall, R. [2004]. The professional response to the aftermath of September 11, 2001, in New York City: Lessons learned from treating victims of the World Trade Center attacks., *in* B. Litz (ed.), *Early intervention for trauma and traumatic loss*, Guilford, New York, NY.

Nowak, K. and Rauh, C. [2005]. The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction, *Journal of Computer-Mediated Communication* **11**(1): 48–55.

Open Geospatial Consortium Inc. [2006]. An Introduction to GeoRSS: A Standards Based Approach for Geo-enabling RSS feeds. Document Number: OGC 06-0503.

O'Reilly, T. and Milstein, S. [2009]. *The Twitter Book*, O'Reilly Media, Inc., Sebastopol, CA.

Ostrow, A. [2008]. Mashable: Is Twitter about to have a big spam problem?, Available at: <`http://mashable.com/2008/03/24/twitter-spam/`>.

Pal, N. and Bezdek, J. [1995]. On cluster validity for the fuzzy c-means model, *IEEE Transactions on Fuzzy Systems* **3**(3): 370–379.

Pang, B. and Lee, L. [2008]. *Opinion Mining and Sentiment Analysis*, Vol. 2 of *Foundation and Trends in Information Retrieval*, now Publishers Inc., Hanover, Boston MA.

Pear Analytics [2009]. Twitter Study August 2009, *Technical report*, Pear Analytics. Available at: <`http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf`>.

Pearson, K. [1895]. Contributions to the mathematical theory of evolution. III. Regression, Heredity, and Panmixia., *Proceedings of the Royal Society of London* **59**(353-358): 69–71.

Pennacchiotti, M. and Popescu, A.-M. [2010]. Detecting controversies in Twitter: a first study, *Proc. NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pp. 31–32.

Petrović, S. [2006]. A comparison between the silhouette index and the Davies-Bouldin Index in labelling IDS clusters, *Proc. NORDSEC 2006*, pp. 53–64.

Petrovic, S., Osborne, M. and Lavrenko, V. [2010]. The Edinburgh Twitter Corpus, *Proc. NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pp. 25–26.

Phelan, O., McCarthy, K. and Smyth, B. [2009]. Using Twitter to recommend real-time topical news, *Proc. RecSys09*, pp. 385–388.

Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L. and Riedl, J. [2007]. Creating, destroying, and restoring value in Wikipedia, *Proc. GROUP '07*, pp. 259–268.

Puniyani, K., Eisenstein, J., Cohen, S. and Xing, E. P. [2010]. Social links from latent topics in microblogs, *Proc. NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pp. 19–20.

Rahman, N. [1968]. *A Course in Theoretical Statistics*, Charles Griffin and Company.

Ramsden, A. [2008]. How and why are people using Twitter: A small group study, *Technical report*, University of Bath.

Rangaswamy, N., Jiwani, S. and Chowdhury, I. R. [2010]. Micro-blogging or GupShup (chatter): Mobile chattering in India, *Proc. CHI 2010 Workshop on Microblogging*.

Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F. and Vespignani, A. [2010]. Characterizing and modeling the dynamics of online popularity, *Phys. Rev. Lett.* **105**: 158701.

Raymond, M. [2010]. How Tweet It Is!: Library acquires entire Twitter archive: Library of Congress Blog, Available at: <`http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/`>.

Relax News [2009]. Current Twitter trends: Google Wave, 'A real wife', Available at: <`http://www.independent.co.uk/news/media/current-twitter-trends-google-wave-a-real-wife-1820222.html`>.

Ritter, A., Cherry, C. and Dolan, B. [2010]. Unsupervised modeling of Twitter conversations, *Proc. NAACL HLT 2010*, pp. 172–180.

Rogers, S., Sedghi, A. and Evans, L. [2011]. UK riots: every verified incident - interactive map, Available at: <`http://www.guardian.co.uk/news/datablog/interactive/2011/aug/09/uk-riots-incident-map`>.

Russell, M. A. [2011a]. *21 Recipes for Mining Twitter*, O'reilly Media.

Russell, M. A. [2011b]. *Mining the Social Web*, O'Reilly Media.

Sakaki, T., Okazaki, M. and Matsuo, Y. [2010]. Earthquake shakes twitter users: real-time event detection by social sensors, *Proc. WWW '10*, pp. 851–860.

Saputra, A. and Leitsinger, M. [2009]. Deadly blasts hit Jakarta hotels, Available at: <`http://edition.cnn.com/2009/WORLD/asiapcf/07/16/indonesia.hotel.explosion/index.html`>.

Sarno, D. [2009]. Twitter creator Jack Dorsey illuminates the site's founding document, Available at: <`http://latimesblogs.latimes.com/technology/2009/02/twitter-creator.html`>.

Schafer, D. L. [2010]. Identifying anomalous users in social networks, Carnegie Mellon University Senior Thesis. Undergraduate Thesis.

Schawbel, D. [2011]. 5 reasons why your online presence will replace your resume in 10 years, Forbes. Available at: `http://www.forbes.com/sites/danschawbel/2011/02/21/5-reasons-why-your-online-presence-will-replace-your-resume-in-10-years/`.

Schrammel, J., Koffel, C. and Tscheligi, M. [2008]. How much do you tell? Information disclosure behavior in different types of online communities, *Proc. 4th International Conference on Communities and Technologies*, pp. 275–284.

Segev, A. and Kantola, J. [2011]. Patent service self organizing maps, *Proc. ITNG 2011.*

Serbanuta, C., Chao, T. and Takazawa, A. [2010]. Save the tweets so you can understand the birds, *Proc. iConference 2010.*

Shamma, D. A., Kennedy, L. and Churchil, E. F. [2009]. Tweet the debates: Understanding community annotation of uncollected sources, *Proc. ACM Multimedia 2009.*

Shamma, D. A., Kennedy, L. and Churchill, E. F. [2010]. Media, conversations and shadows, *Proc. CHI 2010 Workshop on Microblogging.*

Sharifi, B., Hutton, M.-A. and Kalita, J. [2010]. Summarizing microblogs automatically, *Proc. NAACL HLT 2010*, pp. 685–688.

Shaw, W. S. [2009]. Riotous Sydney: Redfern, Macquarie Fields, and (my) Cronulla, *Environment and Planning D: Society and Space* **27**: 425–443.

Simon, M. [2008]. Student 'twitters' his way out of Egyptian jail, CNN Inc. Available at: <`http://www.cnn.com/2008/TECH/04/25/twitter.buck/`>.

Slater, A. S. and Feinman, S. [1985]. Gender and the phonology of north American first names, *Sex Roles* **13**(7–8): 429–440.

Smith, A. and Brenner, J. [2012]. Twitter Use 2012, *Technical report*, Pew Research Center. Available at: <`http://pewinternet.org/~/media/Files/Reports/2012/PIP_Twitter_Use_2012.pdf`>.

Smith, M. A., Ubois, J. and Gross, B. M. [2005]. Forward thinking, *Proc. Conference on Email and Anti-Spam (CEAS 2005).*

Solomon, D. [2008]. How effective are individual lifestyle changes in reducing electricity consumption? - Measuring the impact of Earth Hour, *Technical report*, University of Chicago, Graduate School of Business.

Song, M. [2008]. The new Face(book) of Malaysian cyberfeminist activism, Honors thesis, School of Arts & Social Sciences: Monash University Malaysia.

Spearman, C. [1904]. The proof and measurement of association between two things., *American Journal of Psychology* **15**: 72–101.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M. [2010]. Short text classification in Twitter to improve information filtering, *Proc. SIGIR '10*, pp. 841–842.

Starbird, K., Palen, L., Hughes, A. and Vieweg, S. [2010]. Chatter on The Red: What hazards threat reveals about the social life of microblogged information, *Proc. CSCW 2010.*

Stoica, A., Couronne, T. and Beuscart, J.-S. [2010]. To be a star is not only metaphoric: From popularity to social linkage, *Proc. Fourth International AAAI Conference on Weblogs and Social Media*.

Stonedahl, F., Rand, W. and Wilensky, U. [2010]. Evolving viral marketing strategies, *Proc. GECCO '10*, pp. 1195–1202.

Student [1908]. The probable error of a mean, *Biometrika* **6**(1): 1–25.

Subramanian, S. and March, W. [2010]. Sharing Presence: Can and should your tweets be automated?, *Proc. CHI 2010 Workshop on Microblogging*.

Suh, B., Hong, L., Convertino, G., Chi, E. H. and Bernstein, M. [2010]. Sensemaking with tweeting, *Proc. CHI 2010 Workshop on Microblogging*.

Surowiecki, J. [2005]. *The Wisdom of Crowds*, Abacus, London.

Sutton, J., Palen, L. and Shlovski, I. [2008]. Back-channels on the front lines: Emerging use of social media in the 2007 Southern California Wildfires, *Proc. 2008 ISCRAM Conference*.

Sysomos Inc. [2010]. Twitter Statistics for 2010: An in-depth report at Twitters Growth 2010, compared with 2009, *Technical report*, Sysomos Inc. Available at: <`http://www.sysomos.com/insidetwitter/twitter-stats-2010/`>.

Terdiman, D. [2008]. Twitter Japan launches, with ads, CNET News. Available at: <`http://news.cnet.com/8301-13772_3-9926331-52.html`>.

Terdiman, D. [2009]. Photo of Hudson River plane crash downs TwitPic, CNET News. Available at: <`http://news.cnet.com/8301-1023_3-10143736-93.html`>.

The Associated Press [2009]. Military Weighs Public Outcry on Twitter, Available at: <`http://www.cbsnews.com/stories/2009/08/10/national/main5228836.shtml`>.

The Associated Press [2011]. Tweets-per-second mark set during final, Available at: <`http://espn.go.com/sports/soccer/news/_/id/6779582/women-world-cup-final-breaks-twitter-record`>.

The Nielsen Company [2010]. The State Of Mobile Apps, *Technical report*, The Nielsen Company. Available at: <`http://blog.nielsen.com/nielsenwire/wp-content/uploads/2010/09/NielsenMobileAppsWhitepaper.pdf`>.

The Sunshine Press [2009]. WikiLeaks: 9/11 tragedy pager intercepts, Available at: <`http://911.wikileaks.org/`>.

Thom-Santelli, J., DiMicco, J. M. and Millen, D. R. [2010]. Cross-cultural analysis of status messages within IBM, *Proc. CHI 2010 Workshop on Microblogging*.

Thomas, K., Grier, C., Paxson, V. and Song, D. [2011]. Suspended accounts in retrospect: An analysis of Twitter spam, *Proc. IMC11*.

Tien, J. [2005]. Viewing urban disruptions from a decision informatics perspective, *Journal of Systems Science and Systems Engineering* **14**(3): 257–288.

Tonkin, E. and Tourte, G. [2012]. Twitter, information sharing and the london riots?, *Bulletin of the American Society for Information Science and Technology* **38**(2): 49–57.

Tønnevold, C. [2009]. The Internet in the Paris Riots of 2005, *Digitising the Public Sphere* **16**(1): 87–100.

Torres, P. [2004]. *Visualizing Social Networks: A social network visualization of groups in the online chat community of Habbo Hotel*, Master's thesis, Parsons School of Design.

Tou, J. and González, R. [1974]. *Pattern recognition principles*, Applied mathematics and computation, Addison-Wesley Pub. Co.

Troy, D. [2011]. *Twittervision*, Available at: <http://www.twittervision.com>.

Twitpic Inc. [2009]. Twitpic - share photos on Twitter, Available at: <http://twitpic.com/>.

Twitter Inc. [2009]. *Twitter*, Available at: <http://www.twitter.com>.

Twitter Inc. [2011a]. History of the REST & Search API - Twitter Developers, Available from https://dev.twitter.com/docs/history-rest-search-api.

Twitter Inc. [2011b]. Twitter Blog, Available at: <http://blog.twitter.com/>.

Twitter Inc. [2012a]. Twitter API Documentation, Available at: <http://apiwiki.twitter.com/Twitter-API-Documentation>.

Twitter Inc. [2012b]. Twitter Blog: Twitter turns six, Available at: <http://blog.twitter.com/2012/03/twitter-turns-six.html>.

U.S. Census Bureau [2010]. Genealogy data: Frequently occurring surnames from Census 1990, Available at: <http://www.census.gov/genealogy/www/data/1990surnames/index.html>.

U.S. Social Security Administration [2011]. Baby Name Data, Available at: <http://www.socialsecurity.gov/OACT/babynames>.

van Liere, D. [2010]. How far does a tweet travel? Information brokers in the Twitterverse, *Proc. MSM10.*

Vertesi, J. [2010]. Tweeting spacecraft, *Proc. CHI 2010 Workshop on Microblogging.*

Vieweg, S. and Starbird, K. [2010]. Microblogging in mass emergency, *Proc. CHI 2010 Workshop on Microblogging.*

Wagner, C. and Strohmaier, M. [2010]. The wisdom in tweetonomies: acquiring latent conceptual structures from social awareness streams, *Proc. SEMSEARCH '10*, pp. 6:1–6:10.

Ward, Jr., J. H. [1963]. Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association* **58**(301): 236–244.

Warden, P. [2011]. Data Science Toolkit, Available at: <`http://www.datasciencetoolkit.org/`>.

Warden, P. [2012]. OpenHeatMap, Available from `http://www.openheatmap.com/`.

Wasik, B. [2009]. *And Then There's This: How Stories Live and Die in Viral Culture*, Penguin Group (USA), New York, NY.

Wasow, O. and Baron, A. [2010]. Can tweets kill a movie? An empirical evaluation of the Bruno Effect, *Proc. CHI 2010 Workshop on Microblogging.*

Watne, T. and Cheong, M. [2012]. #Socialisation Agency or @Branding? How 'Mass Brewers' and 'Craft Brewers' communicate with consumers through Twitter, *Submitted for consideration to EMAC 2013.*

Westman, S. and Freund, L. [2010]. Information interaction in 140 characters or less: Genres on Twitter, *Proc. IIiX 2010.*

Wigand, F. D. L. [2010]. Twitter takes wing in government: diffusion, roles, and management, *Proc. DG.O '10*, pp. 66–71.

William-Ross, L. [2009]. LAist: Lights out, Los Angeles: Earth Hour is tonight., Available at: <`http://laist.com/2009/03/28/lights_out_los_angeles_earth_hour_i.php`>.

Wilson, D. W. [2008]. Monitoring technology trends with podcasts, RSS and Twitter, *Library Hi Tech News* **25**(10): 8–12.

Wu, W., Zhang, B. and Ostendorf, M. [2010]. Automatic generation of personalized annotation tags for Twitter users, *Proc. HLT '10*, pp. 689–692.

Xu, K. and Farkas, D. K. [2008]. Blogging as a rhetorical act, *Proc. 73rd Association for Business Communication Annual Convention.*

Yamazaki, Y. and Kumasaka, K. [2010]. Comparative analysis of the people's character of Japanese prefectures based on SNS, *Proc. ICIS 2010.*

Yao, K. B. [2006]. *A comparison of clustering methods - for unsupervised anomaly detection in network traffic*, Master's thesis, Department of Computer Science, University of Copenhagen (DIKU).

Yao, Z., Eklund, T. and Back, B. [2010]. Using SOM-Ward clustering and predictive analytics for conducting customer segmentation, *Proc. 2010 ICDM Workshops.*

Yoshida, M., Inui, T. and Yamamoto, M. [2010]. Analysis of tweets including URLs on Twitter (translated), *Proc. DEIM Forum 2010.*

Zarrella, D. [2009]. State of the Twittersphere June 2009, *Technical report*, HubSpot Inbound Internet Marketing. Available at: <`http://blog.hubspot.com/Portals/249/sotwitter09.pdf`>.

Zhang, J., Qu, Y. and Hansen, D. [2010]. Modeling user acceptance of internal microblogging at work, *Proc. CHI 2010 Workshop on Microblogging*.

Zhao, D. and Rosson, M. [2009]. How and why people Twitter: the role that micro-blogging plays in informal communication at work, *Proc. GROUP'04*, pp. 243–252.