Information Propagation in Twitter

Alexandros Toumazis

University of Cambridge
Computer Laboratory
Fitzwilliam College

June 2010

MPhil in Advanced Computer Science project dissertation

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the regulation length of 15 000 words, including tables and footnotes.

Information Propagation in Twitter

Alexandros Toumazis

Abstract

Twitter is a large and popular social network which has grown, and changed, extremely rapidly over the past year, and has become an important source of news and information for many users. This growth and change has made Twitter not only very influential, but also a unique blend between a social network and a broadcasting medium. Understanding how information propagates within this network can give us insights into what types information spread successfully, and which users are more successful in spreading information, which has implications for marketing and news distribution. This project aimed to analyze and characterize the spread of information in Twitter, and discover factors which affected this spread.

A large set of Twitter data was obtained and analyzed using established metrics to find factors which led to significant differences. Two such factors were found: language and topic. People in different language groups differ in the extent that they regard Twitter as a social network or news source, and this affects their activities. These differences could be due to differing ways in which people interact with Twitter, cultural differences in the way they use the Internet, or the presence or absence of competing social networks. These potential explanations are backed up by reference to previous research and new data showing significant inter-language differences in the interfaces people use to interact with Twitter. Similarly, conversation about different topics leads to measurable differences in these metrics.

Acknowledgments

Thanks to Eiko Yoneki for guidance and countless ideas, Jon Crowcroft for suggesting tracking the General Election, and Narseo Vallina-Rodriguez and Damien Fay for useful suggestions and proofreading.

Contents

1 Introduction **12**

1.1 Structure 13

2 Background and Related Work **15**

2.1 Twitter 15

2.1.1 A Brief History 15

2.1.2 Terminology 16

2.2 Related Work 17

3 Data Collection and Methodology **19**

3.1 Data Collection 19

3.1.1 Trending Topics 20

3.1.2 Random Sample 21

3.1.3 Election Data 21

3.2 Analysis 22

3.2.1 Basic Metrics 22

3.2.2 Language Determination 22

3.2.3 Retweet Tree/Graph Construction 23

3.2.4 Election 24

4 Characterizing Tweets and Users **26**

4.1 Characterizing Tweets 26

4.1.1 Mentions, Retweets, Hashtags and URLs 26

4.2 Characterizing Users 27

4.2.1 Activity 27

4.2.2 Influence 28

5 Information Spread **31**

5.1 Characterizing Information Spread 31

5.1.1 Retweet Trees 31

5.1.2 Retweet Graphs 32

5.1.3 Topics 33

5.2 Language 33

5.2.1 Language Popularity 34

5.2.2 Characterizing Languages 34

5.2.3 Individual Measures 34

5.2.4 Combined Measures 36

5.2.5 Potential Explanations 36

5.2.6 Future Directions 37

5.3 Topic 38

5.3.1 Topic Categorization 38

6 Case Study: UK General Election **44**

6.1 Timeline 44

6.2 Temporal Analysis 45

6.3 Party Affiliation 47

6.4 Parties and Buzzwords 49

6.5 Sentiment Analysis 50

6.6 Summary 51

7 Conclusions & Future Directions **52**

7.1 Conclusions 52

7.2 Future Directions 53

7.2.1 Using Location Information 53

7.2.2 Multicasting, Channeling, and Aggregating 53

A Terms and Users Used to Acquire Election Data **57**

B Users with known party affiliation **59**

C Election Graphs **60**

D Trending Topics **65**

List of Tables

3.1 Top 10 trigrams in English and Portuguese reference sets 23

3.2 Similariy measures between example tweets and reference sets 23

4.1 Metrics of a random sample of tweets compared 26

4.2 Metrics of a random sample of retweets compared 27

4.3 Top 10 users using three influence measures 30

5.1 Comparison of language distribution results 34

6.1 Party affiliation assigned, based on seeds given in B.1, from a 157,764 tweet sample from 05:43 7/5/2010 to 23:17 11/5/201 49

A.1 Terms used to filter election data 58

B.1 Seeds for party affiliation propagation 59

D.1 Trending topics with over 10,000 tweets (1) 65

D.2 Trending topics with over 10,000 tweets (2) 66

D.3 Trending topics with over 10,000 tweets (3) 67

D.4 Trending topics with over 10,000 tweets (4) 68

D.5 Trending topics with over 10,000 tweets (5) 69

D.6 Trending topics with over 10,000 tweets (6) 70

D.7 Trending topics with over 10,000 tweets (7) 71

D.8 Trending topics with over 10,000 tweets (8) 72

D.9 Trending topics with over 10,000 tweets (9) 73

D.10 Trending topics with over 10,000 tweets (10) 74

D.11 Trending topics with over 10,000 tweets (11) 75

D.12 Trending topics with over 10,000 tweets (12) 76

D.13 Trending topics with over 10,000 tweets (13) 77

List of Figures

3.1 Tweet class used for data gathered from all APIs 20
3.2 User class used for data gathered from Streaming API 20
3.3 Tweet A: Example of an English tweet 22
3.4 Tweet B: Example of a Portuguese tweet 22
3.5 Example of a correctly categorized positive tweet about the Labour party. . 24
3.6 Example of a correctly categorized negative tweet about the Conservative party 25
3.7 Example of an incorrectly categorized tweet: it was categorized as positive 25

4.1 Tweet/user distribution within a sampled set of tweets 28
4.2 Average retweet and mention counts against number of followers (For a sample of 1,360,000 random tweets) 28
4.3 Average retweet count against mention count (For a sample of 1,360,000 random tweets) 29

5.1 Example of a retweet tree (from trending topic `#RIPAlejandraJonas`) . . . 32
5.2 Retweet tree size/frequency for retweet trees in the topic 'Haiti' 33
5.3 Example of a retweet graph 39
5.4 Mention rate and frequency for the 12 most popular languages 40
5.5 Retweet rate for the 12 most popular languages 40
5.6 Hashtag rate and frequency for the 12 most popular languages 41
5.7 URL rate for the 12 most popular languages 41
5.8 Combined (normalized) metrics for the five most popular languages 42
5.9 Tweet source distribution for the four most popular languages 42
5.10 Differences between five popular topics 43

5.11 Scatter graph showing manually categorized topics 43

6.1 Party mentions during the week around the election 46

6.2 Party leader mentions during the week around the election 46

6.3 Party leader mentions and the FTSE 100 Index during the day after around the election 47

6.4 Graph colored by party affiliation from a 157,764 tweet sample from 05:43 7/5/2010 to 23:17 11/5/2010 48

6.5 Pre-election coincidence of keywords and party leaders 50

6.6 Pre-election coincidence of american politicians and party leaders 50

6.7 Twitter users positive/negative views towards the parties in the day before the election 51

C.1 Party leader mentions and the FTSE 100 Index during the day after around the election 60

C.2 Party leader mentions during the week around the election 61

C.3 Party mentions during the week around the election 62

C.4 Graph colored by party affiliation from a 157,764 tweet sample from 05:43 7/5/2010 to 23:17 11/5/2010 63

C.5 Graph colored by party affiliation from a 157,764 tweet sample from 05:43 7/5/2010 to 23:17 11/5/2010 64

Chapter 1

Introduction

Twitter is an interesting and novel blend between a traditional social network, which usually consists of bidirectional friendship links, and a broadcast medium, where users can subscribe to interesting feeds or view data by topic. In this project I show that characteristics of Twitter user activity depend on the topic being discussed and the language being spoken. In addition, I examine the changing nature of the Twitter landscape, most significantly how its shift since launch from a SMS-based[1] service to a web service has affected the way people use it.

Twitter is large, important and rapidly changing: 87% of Americans have heard of it (compared with 26% a year ago), and its userbase has more than quadrupled to over 75 million over the past year. Celebrities and conventional news sources have embraced it, both by setting up their own user accounts and by integrating Twitter feeds into their products — for example, Major League Baseball now provides a timeline of game-related tweets within its online live game video service. Twitter is also very international: according to some measures, less than half the tweets posted on Twitter are in English, and, because of its mobile-friendly nature, it has become popular in many developing countries where at-home internet connections are less common. As Twitter becomes less like a social network and more like an information source, it will open up to marketing, as users tend to regard advertising in their information feeds as normal, while they push back against intrusive advertisements in social networks.

Understanding how information propagates in Twitter is essential for creating a strategy to effectively spread it. Social networks have various different mechanisms which can be used for spreading information, and the effectiveness of each one depends on a host of factors. Work has been done empirically analyzing data from Twitter as a whole; however, little work has looked into what factors can cause significant changes in these measurements. Following a qualitative examination of the characteristics of information

[1]SMS: Short Message Service, a standard text message service provided on most mobile phone network; message length is limited to 140 characters in a single message

spread in Twitter, I acquired and analyzed empirical metrics, such as the frequency of token denoting social links and the frequency of links to external information, to find factors that affected them. The data I used to do this was collected using Twitter's public APIs, and is described in Chapter 3. I also quantitively compared the data gathered with previous work, leading into insights into the way Twitter is changing.

I found that different regions — defined by language — use Twitter in very different ways, with some regions treating it as a social network but others using it as a global or topic-keyed information source. This affects how information spreads and therefore the success of a particular strategy or user in disseminating a particular message. Also, similar differences are observed between trending topics, with similar implications. These inter-topic differences may also make automated topic categorization feasible. In addition, a comparison of general results to related work and reference to the development of Twitter shows its general shift from a social network to a more broadcast-like medium.

Knowledge of how Twitter is being used within a certain topic can lead to more refined marketing techniques: in a topic with broadcast-like behavior, interesting or engaging global tweets will grab the populations attention, while for more social topics, a slow build-up of social and trust links may be necessary before users can be effectively drawn in. This information could also be used to provide effective topic recommendations, as the type of topic is important when deciding how to recommend it: a user is likely to want community-based topic recommendations drawn from his social links, but broadcast-based topic recommendations drawn from his global interests. Understanding of how Twitter use differs across regions is vital for designing marketing campaigns in different regions. There are also implications for caching strategies for media linked to on Twitter based on predictions on how the link will spread. Understanding how use of Twitter has changed and is changing can enable us to make predictions as to how it will change in the future, which again has obvious implications for the design of future Twitter client, services or marketing campaigns.

Finally, I also present a topic case study of the UK general election, including a temporal analysis showing how Twitter users reacted to breaking news and events, sentiment and keyword analysis of election-related tweets, and a method for determining users' party affiliation. This chapter presents a concrete example of how users simultaneously use Twitter as a social network, in following and retweeting users they agree with, and as a broadcast medium, using hashtags to proselytize their views to other users tracking the election, as well as offering several useful tools for marketing and brand analysis.

1.1 Structure

The structure of the remainder of this dissertation is as follows: Chapter 2 explains how Twitter works, its history and terminology, and describes related work on social networks

and Twitter. Chapter 3 describes the implementation of the data collection and analysis for this project. Chapter 4 presents basic metrics and characteristics of tweets and users, and Chapter 5 describes information spread and characterizes the variations between conversation in different languages and topics. Chapter 6 presents a topic case study of the 2010 UK General Election. Chapter 7 contains conclusions and potential future directions.

Chapter 2

Background and Related Work

This chapter explains the way Twitter works, provides a history of its development and describes the terminology used in this dissertation, before going on to describe related work about Twitter and social networks in general.

2.1 Twitter

2.1.1 A Brief History

Twitter launched in 2006 as a mainly SMS-based service aimed at allowing users to broadcast messages to their friends from anywhere. As users would receive a text message for each of their followees tweets, users tended to follow a relatively small number of people, and user activity in general was relatively low. Since then, Twitter use has shifted to a web- and API-based services, which has significantly changed the demographics and use patterns of users. The Twitter user population exploded in 2009, increasing from under 10 million to over 75 million users, but most metrics for user engagement fell: 80% of users at the end of 2009 had tweeted a total of under 10 times, and average follower numbers have fallen[Moo10]. This is probably due to the large amount of media attention Twitter received that year, due in part to the Iranian election, bringing many new, and different, users to the service. In 2009, Twitter also changed the question presented to tweeting users from "What are you doing?" to "What's happening?" and began providing a search facility and a list of trending topics. These changes were a reaction to this change in use patterns, and show a shift in tone from asking users to share personal information not necessarily of general interest to soliciting more newsworthy tweets. However, these changes have not been complete or global — different regions use Twitter in different ways and from different devices, and appear on a spectrum between the original, friend-centric model, and the new global information and news model. These differences also present themselves at the level of individual topics, although in this case they are not necessarily

due to demographic or usage changes, but inherent to the type of discussion present in different categories of topic.

2.1.2 Terminology

Twitter has lots of specific jargon, and Twitter research uses even more. Readers familiar with Twitter can skip this section, as it just offers a quick overview of the terminology used in Twitter and in this thesis. A *tweet* is the basic unit on twitter, and is a message no longer than 140 characters posted by a specific *user* at a certain time. By convention and for clarity, I will prefix user's name with the '@' symbol. A user can *follow* other users — this creates a one-way relationship from the user to his followee. A user's default view is of their twitter *feed*, which contains, ordered by recency, her tweets and the tweets of her followees. In addition, a user can search for tweets matching a supplied string; the search phrase is called a *topic*. The most popular (in terms of posting, not searching) topics for various timeframes and levels of geographic locality are displayed on the Twitter website, and are referred to as *trending topics*. There are various pieces of extra information that can be present in a tweet; these developed through convention and eventually became institutionalized in the Twitter API:

- *Mentions* — a @ followed by a user's screen-name in a tweet will make that tweet appear on the referenced user's feed, for example "`hey @alice, what's up?`"
- *Retweets* — By convention, when reposting another users tweet, users prefix the tweet with 'RT ' and the original tweeter's screen-name. This has become part of the Twitter API: a user can click an icon on a tweet in their feed and this prefix will be silently appended and the tweet reposted, for example "`RT @bob: <amusing anecdote>`", and
- *Hashtags* — a # followed by a string is used to explicitly define a topic, for example "`oil is bad for birds #oilspill`".

In addition, there are some pieces of meta-data that can be associated with a tweet:

- *Creation date/time* — Self-explanatory; present in every tweet and converted to reader's local timezone,
- *Source* — Provides information on which service/application the tweet was posted from. Third-party services and applications can define their own strings, and Twitter itself uses 'API', 'txt' and 'web', and
- *Location* — Provides per-tweet location information provided by the user's client software. This is a recent addition to the API and currently not widely used.

When used in the main text, tweet contents, user names, hashtags and topic will be written in `this typeface`.

2.2 Related Work

There exists a wide range of related work, both about Twitter and social networks in general. This section focusses on work directly related to this project, including studies of information propagation in Twitter and other social networks, measuring influence and using geographic and cultural information to improve performance or the user experience. Over half the papers referenced here were published in the past year and a half, which shows how quickly research is moving in this field.

Cha et al. [CMG09] study information propagation in Flickr. The paper begins with similar aims to this dissertation, but diverges due to the radically different nature of the way users user Flickr. Flickr does have a similar one-way friend relation, leading to some of the same effects as observed in Twitter — namely a small number of very, very popular users whose content is widely spread. However, information cascades are common in Flickr, unlike Twitter. Gruhl et al. [GGLNT04] look at information diffusion between blogs, a space which has similar internal/external content divisions as Twitter, dividing topics between short-lived spikes, usually triggered by external events, and long-lived chatter driven mostly by internal content and comments.

Kwak et al. [KLPM10] cover some of the same ground as this project, presenting similar basic measures, but the authors focus on using temporal characteristics to study topics instead of the mainly time-independent measures used in this project. They also use PageRank to rank users, which was not possible for this project due to not having access to the Twitter social graph. Boyd et al. [BGL10] describe Twitter conventions and results of qualitative user studies of Twitter, and provide useful information on how and why people retweet, as well as metrics on a sample of random tweets taken in early 2009, which I compare with my results in Section 4.1.1. Java et al. [JSFT07] also look at the reasons why people tweet, and analyses the Twitter social network. The paper also provides some basic geographic and language-based analysis, and is useful in that it provides details and a snapshot of Twitter as it was three years ago, when it was still mainly a social network used via mobile phones.

Daly [Dal09] discusses deriving user reputation measures (similar to the 'influence' measures usually used for Twitter) and dynamically using them to rank data in real-time. Cha et al. [CHBG10] examine the challenge of measuring user influence in Twitter, specifically the misleading nature of follower count, and propose new metrics for measuring user influence. In Chapter 4 I describe their findings and compare them with my results. Huberman et al. [HRW08] also discuss the unsuitability of follower count as a useful metric,

as well as defining a new 'friend' link between users who mention each other and showing that it correlates with user activity much better than follower count.

Benevenuto et al. [BRCA09] analyze the Orkut social network and the spread of information among its users, discusses analyzing geographic spread, and show that most social network activity does not result in any visible signs — i.e. it is passive browsing or searching. Sastry et al. [SYC09] analyze geographic spread of YouTube video views based on social links in Facebook and provide an example of a possible application requiring being able to characterize information spread. Geographic spread in a strict sense is not covered in this dissertation, except for Section 7.2.1, but the applications of knowing in which languages content is being posted, and to which languages it is likely to spread is equivalent, although coarser.

Chau et al. [CCM+02] discuss how cultural differences between regions can lead to people using the Internet in different ways, and provide a user study showing American consumers feeling more comfortable with search-based interfaces compared with consumers in Hong Kong, who were more comfortable with community-based systems. This is a potential explanation for the differences observed in Section 5.2

Zhao and Rosson [ZR09] discuss how people use Twitter within businesses, and regard Twitter as a 'mirco-blogging' service, a fundamental new type of social network. Reference is made to users wanting easier ways to limit the reach of their tweets and see more easily what categories (such as 'co-worker' or 'friend') other users, and to users using their followees as a sort of filter, bringing to their attention only information they are interested in; this concept is also discussed in Section 7.2.2. This contrasts with automated content recommendation systems, such as the complex one described in [CNN+10], which use a multitude of global and local factors to attempt to recommend useful content.

Jansen et al. [JZSC09] describe the potential of Twitter for real-time marketing and brand analysis, and perform sentiment analysis on brand-related tweets. This ties in with my work in Section 6.7, and the rest of Chapter 6 also has significant applications in brand analysis.

Chapter 3

Data Collection and Methodology

This chapter describes the data gathered, along with the methodology and reasoning used to acquire it and analyze it, and the metrics used to characterize it. Deeper analysis and specific conclusions reached are described in Chapters 4, 5 and 6.

3.1 Data Collection

To acquire the data for my project, I used three APIs provided by Twitter:

- Search API: Returns the most recent tweets containing a specific phrase, up to a maximum of 1,500 tweets. I used this API to scrape tweets from trending topics,
- REST API: This API provides various forms of meta-data and is rate-limited to 150 queries/hour. I used this API to acquire social graph information for preliminary work, and for acquiring the currently trending topics, and
- Streaming API: Returns real-time data[1], either a random sample of all tweets or all tweets matching specified topic and/or username filters.

To store the gathered data I used two Python classes, one to store user details returned by the Streaming API (Figure 3.2), and one to store tweets returned by all three APIs (Figure 3.1). The tweet class contains the basic information returned by the API, and variables to hold graph and language information. The user class saves users' basic social graph metrics and location information.

[1] In practice, due to the large volume of data this feed swiftly fell behind when streaming a sample of all tweets.

```
class Tweet:
  def __init__(self, idnum, screen_name, datetime, text, source='web',
      language=None):
    self.id=idnum
    self.screen_name=screen_name
    self.datetime=datetime
    self.text=text
    self.source=source
    self.rt=[]
    self.parents=[]
    self.language=language
  def printOneLine(self):
    ...
```

Figure 3.1: Tweet class used for data gathered from all APIs

```
class User:
    def __init__(self, screen_name, idnum, friends_count, followers_count,
        loc, lat=None, lon=None, utc_offset=None):
    self.screen_name = screen_name
    self.id = idnum
    self.friends_count=friends_count
    self.followers_count=followers_count
    self.location=loc
    self.lat=lat
    self.lon=lon
    self.utc_offset=utc_offset
  def printBasic(self):
    ....
```

Figure 3.2: User class used for data gathered from Streaming API

3.1.1 Trending Topics

To acquire tweets from trending topics, I maintained a list of trending topics using the `topics` method of the Twitter REST API, which provides the current top ten trending topics. Every 5 minutes I updated this list by requesting the current trending topics and adding them to the list or updating the last seen time where appropriate, and used the Search API to scrape all currently 'fresh' topics, 'fresh' being defined as having appeared on the trending topic list no more than 12 hours ago. To avoid duplication of tweets and reduce bandwidth demands, the program keeps track of the most recently seen tweet from each topic and only goes back that far in searching for new tweets.

The final data set consists of all topic for which over 10,000 tweets were collected: this results in a set of 419 topics with an average of 36,489 tweets each.

3.1.2 Random Sample

For a random sample of tweets, I used the `statuses/sample` method of the Streaming API, which is documented as containing a random 5% of all tweets. Several samples were taken over different time periods over a 10 day interval.

The data set collected consists of 9 contiguous blocks of sampled data scraped between 8/4–17/4/2010 and containing a total of 3,833,000 tweets.

3.1.3 Election Data

The `statuses/filter` method of the Streaming API was used with a set of hashtags, topics and users related to the election (Appendix A contains the complete list of terms used).

- *Hashtags:* Several popular election hashtags were included in the filter, and this list was added to manually over the sampling period as new relevant hashtags appeared. Twenty-four were tracked in all, from election-specific topics like `#GE2010` and `#hangem` to party and politician tags like `#imvotinglabour` and `#dcameron`.
- *Topics:* To track general conversation about the election, several topics such as "`Labour`", "`Tory`", "`Clegg`" were used. Only topics that unambiguously were related to the election were included[2], so topics like "`Brown`" and "`Cameron`" were not included. Eight such topics were tracked[3].
- *Users:* The only users included were ones who exclusively tweeted about the general election, to ensure irrelevant tweets were not included in the sample. The API included, in addition to all tweets from these users, all tweets mentioning or retweeting them. Over the first few days of the study, I identified the most active, retweeted and mentioned users in the sample to date and, if they were exclusively posting about the election, added them to the filter. The users were official party accounts and politicians, such as `@Nick_Clegg` or `@labourparty`, and a total of 11 users were tracked.

The final data set consists of 1,108,562 tweets collected between 03:04 on 04/05/2010 to 23:17 12/05/2010.

[2] "`Labour`" was a slight exception, but it overwhelmingly referred to the election (and the British spelling eliminated any American uses)

[3] The Twitter API makes no distinction between topics and hashtags; the topic "`Clegg`" and hashtag `#clegg` are equivalent. Therefore, there is some overlap, but I've chosen to distinguish the two categories, as some words were used only as hashtags while others were used mainly without the hash.

3.2 Analysis

3.2.1 Basic Metrics

To determine mention, retweet, URL and hashtag rates I analyzed tweets and categorized them into one or more categories based on simple textual properties:

- A tweet containing one of the strings `RT @`, `RT@` or `via @`(case insensitive) is a retweet.
- A tweet containing a `@` outside one of the string described above contains a mention.
- A tweet containing a `#` contains a hashtag.
- A tweet containing `http://` contains a URL.

Rates and frequencies can then be simply calculated.

3.2.2 Language Determination

I used the random sample of tweets to look at different languages popularity in Twitter. The tweets were assigned languages using a trigram-based classifier, based on [CT94], trained on a random set of $\approx 60,000$ tweets selected from the sample and categorized using Google's language API. From this sample, I built a trigram-based classifier for the 16 most popular languages. This classifier uses reference trigram frequency tables constructed from these reference tweets to determine the language of a tweet by comparing similarity measures between the tweet's trigram frequencies and these reference tables. A tweet is classified as being in a certain language if its similarity measure with that language is the highest, and there is at least a difference of 0.05 between this measure and the next highest language's similarity measure. This threshold was chosen because it provided a good rate of positive identification while keeping false positives to a minimum.

Table 3.1 shows the most common trigrams for English and Portuguese.

@omgthatssotrue: `I hate it when my favorite song comes on the radio in the car, and someone puts it down to talk. (Fan Idea) #omgthatssotrue`

Figure 3.3: Tweet A: Example of an English tweet

@LucianCanito: `Essa vai para toda galera que Trabalha o dia todo e no final do dia vai pro PC "Quem comeca com sono termina domino."`

Figure 3.4: Tweet B: Example of a Portuguese tweet

Rank	English	Portuguese
1	_th	_de
2	the	de_
3	ing	do_
4	he_	que
5	_to	qu_
6	ng_	_co
7	to_	ue_
8	nd_	as_
9	_I_	em_
10	_an	ra_

Table 3.1: Top 10 trigrams in English and Portuguese reference sets

The similarity measure is calculated by multiplying the corresponding frequencies in the two trigram tables, and is normalized by dividing by the product of half of each tables self-similarity. This is shown in Equation 3.1, where M and N are trigram frequency tables, i and j are trigrams, and $f(i, M)$ is trigram i's frequency in table M.

$$S_{M,N} = \frac{\sum_{i \in M} f(i, M) f(i, N)}{\frac{1}{4} \sum_{i \in M} (f(i, M)^2) \sum_{i \in N} (f(i, N)^2)} \tag{3.1}$$

Table 3.2 shows the results of calculating the pairwise similarity between two example tweets, 3.3 and 3.4, and the reference sets for English and Portuguese. Despite the very small number of trigrams in the tweets — there were no trigrams with a frequency over 3 in either tweet — and the low similarity between the tweets and the much larger language samples, a clear, correct identification of each tweet is made.

	English	Portugese
Tweet A	0.410	0.130
Tweet B	0.121	0.390

Table 3.2: Similariy measures between example tweets and reference sets

As the classifier works far more precisely on non-Western languages, due to their trivially distinguishable character sets, for language popularity estimation the "unsure" tweets are divided between Western languages in the same proportion as the positively identified Western language tweets. For all other results, only positively identified tweets were used.

3.2.3 Retweet Tree/Graph Construction

The method I used to find retweet trees within a sample of tweets was to, for each tweet, first determine the users it retweets (if any) by finding strings of the form "`RT @user`",

"`RT@user`" or "`via @user`", and then, for each of these parent users, find all older tweets from the user and select the best match with the current tweet, if one exists. The best match was selected based on scoring each potential parent by looking at which fraction of words in the two tweets were the same, weighting each word according to its length. Because the tweet samples being used were obviously limited to a certain timespan, there was a significant chance of false positives when the real parent tweet was not included in the sample.

Building a retweet graph for a sample of tweets is considerably simpler. They could be built by combining retweet trees for each tweet in the sample, but this is very inefficient ($O(n^2)$) and unacceptably slow for large samples. Instead, as we only care about who retweeted whom, not which specific tweet was retweeted, all that needs to be done is scan each tweet and add an edge for each detected parent (i.e. detect strings of the form "`RT @user`"). This requires only one pass through the sample and avoids the scoring/missing parent problems mentioned above.

3.2.4 Election

Word usage

I analyzed word usage for selected words in conjunction with party or politician names. This two-level analysis allows comparison of rates of keywords use about different parties.

Sentiment

To determine whether twitter activity related to a political party or leader was positive or negative, I employed some very basic sentiment analysis. For each party, I examined all tweets which contained the word 'vote' and the party name, then divided them into positive, negative and ambivalent based on the presence of phrases such as 'don't' or 'out' (for negative) or 'why' or '?' (for ambivalent). To stop spamming users from having undue influence, I examined a maximum of one tweet per user. Although this method is very prone to miscategorization, over a large sample set it provides useful results (and a more refined method would still have problems, sarcasm is fairly common and quite hard to detect).

```
I wish I wasn't apathetic before and had bothered to register and vote
                labour.  But whats done is done now.
```

Figure 3.5: Example of a correctly categorized positive tweet about the Labour party.

```
Dress in black if Tories form a government.  RT if you didn't vote Tory
#UKMOURNS
```

Figure 3.6: Example of a correctly categorized negative tweet about the Conservative party

```
I heard only virgins vote Tory.
```

Figure 3.7: Example of an incorrectly categorized tweet: it was categorized as positive

Party Affiliation

Beginning with a seed set of users with known political affiliation, any user who retweeted any of these was categorized as a potential supporter of that party. Then this step was repeated to categorize users who had retweeted these newly categorized users. This was repeated until the state of the users stabilized, or a preselected iteration limit was reached. Then, any users who were only categorized as supporters of a single party were confirmed in this affiliation, while users who were categorized into two or more potential parties were left unaffiliated.

Temporal analysis

I continuously scraped election-related tweets for a one week period around the election. As tweets are date-stamped, this allows easy extraction of minute-by-minute activity levels (in the form of tweets/minute), as well as allowing filtering and comparisons between activity levels for different topics within the election data set.

Chapter 4

Characterizing Tweets and Users

This chapter discusses the characteristics of the general Twitter user population and of the tweets they produce. Using a random sample of tweets, the metrics described in the previous two chapters are derived and examine how the influence of individual users can be usefully characterized. My results are compared with previous work, and the differences discussed.

4.1 Characterizing Tweets

4.1.1 Mentions, Retweets, Hashtags and URLs

Table 4.1 shows the proportion of tweets that contain at least one retweet, mention, URL or hashtag and, for comparison, the corresponding figures from [BGL10], which uses a sample of tweets taken over the period 26/1/09–6/13/09. Table 4.2[1] shows similar figures for a sample of retweets. My sample consisted of 2,280,000 tweets, of which 319,695 were retweets, collected over a 6-day period using the Twitter Streaming API.

Language	9/4/10–15/4/10	26/1/09–13/6/09 (From [BGL10])
Mention	42%	36%
Retweet	14%	3%
Hashtag	13%	3%
URL	20%	22%
Tweets in sample	2,280,000	720,000

Table 4.1: Metrics of a random sample of tweets compared

[1] 'Enclosed retweet' in the table refers to tweets of the form "`RT @A RT @B ...`" resulting from multiple rewteets

Language	9/4/10–15/4/10	20/4/09–13/6/09 (From [BGL10])
Enclosed Retweet	20%	11%
Hashtag	23%	18%
URL	25%	52%
Tweets in sample	319,695	203,371

Table 4.2: Metrics of a random sample of retweets compared

The results I obtained are significantly different from those in [BGL10], despite similar data sets:

- Both sets of results show an increase in hashtag and URL usage in the sample of retweeted tweets compared with the general sample. However, my results show far less of an increase — in the case of web links my results show a 25% increase, while in [BGL10] the retweet sample exhibits an 136% increase.
- My results show a much higher rate of hashtags(133% increase), retweets(366% increase), and nested retweets (82% increase).

These differences could be due to Twitter adding support for automatic retweeting to its website and API, but could also be indicative of the changing uses and demography of Twitter: for example, the fall in the URL rate in retweets could be due to Twitter's growing popularity and tweet volume causing more and more content to appear within the network rather than outside on the web, and the rise in retweet and hashtag rates could be due to the continuing shift from SMS/mobile use to PC/smartphone use which, with its more powerful clients, allows both easier retweeting and use of hashtags and easier viewing of global topics.

4.2 Characterizing Users

4.2.1 Activity

Active users account for a large proportion of tweets: In a sample of 2.28 million random tweets taken over a 24-hour period, 35% of users present in the sample were responsible for 65% of the tweets. Figure 4.1 shows the distribution of activity among users, which follows a power law. These figures only take into account users who tweeted during this period; by most estimates a large majority of users tweet rarely or never[Moo10].

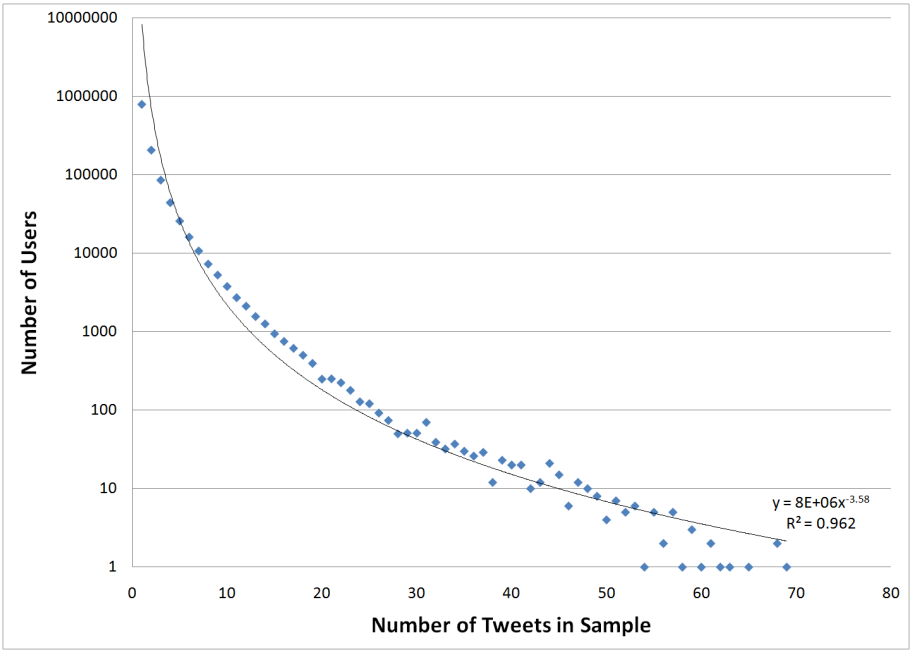


Figure 4.1: Tweet/user distribution within a sampled set of tweets

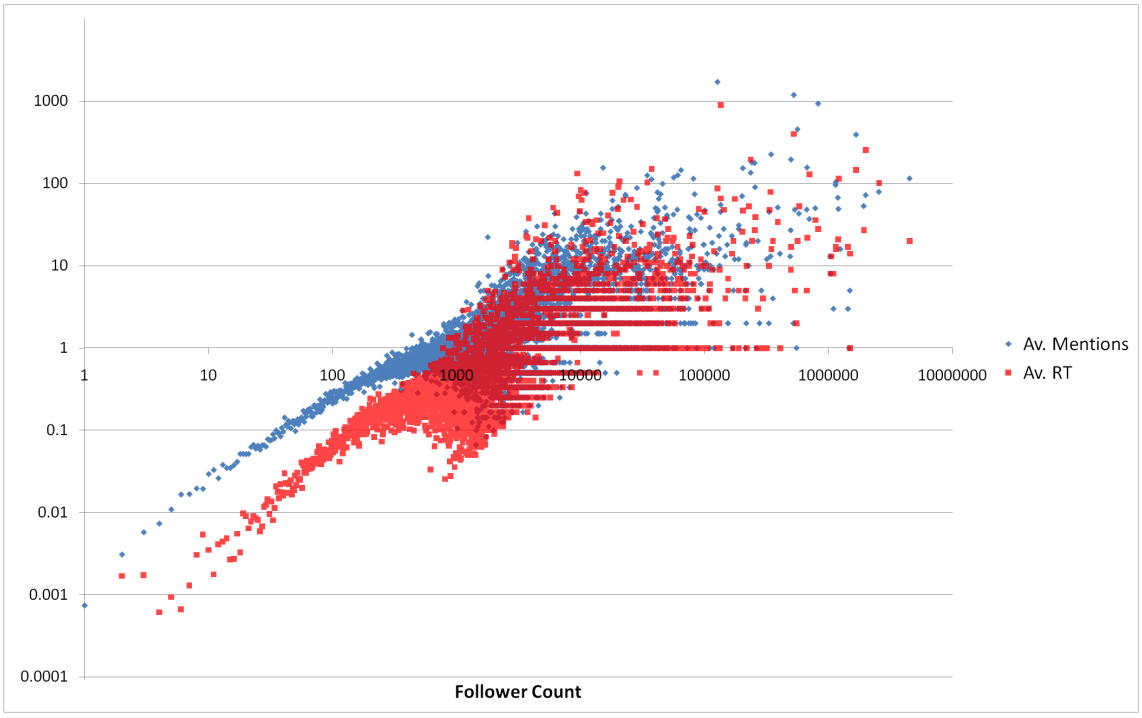


Figure 4.2: Average retweet and mention counts against number of followers (For a sample of 1,360,000 random tweets)

4.2.2 Influence

Measuring a user's influence in Twitter is not straightforward. The simplest measure is follower count (indegree), which has the advantage of being explicit and fairly slow to change. However, [CHBG10] argues that this measure does not give a complete picture,

and is not related to two other proposed measures, namely mention and retweet count. These are important as mention count shows both how engaged in two-way conversation a user is and also how often other users talk about her, and retweet count shows how far information posted by the user is spread, which is the most useful metric for many purposes. [CHBG10] argues that retweets are driven by the content of a tweet and mentions are driven by the name recognition of the user. Given this, we can expect that mentions will in fact be somewhat correlated with follower count, as more recognizable users tend to have more followers, and, to a lesser extent, retweets will also be somewhat correlated with follower count, as more followers leads to more potential retweeters[2]. Figures 4.2 and 4.3 show the correlations between these three measures; below a threshold of about 10000 followers or 1 average mention in the sample set, the measures are related fairly well by a power law; above this threshold the relation breaks down.

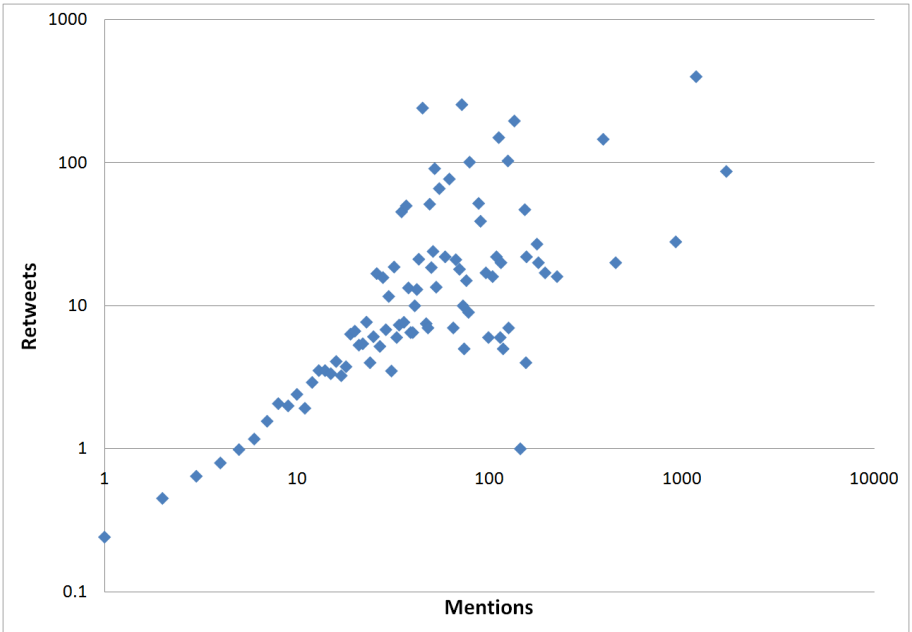


Figure 4.3: Average retweet count against mention count (For a sample of 1,360,000 random tweets)

Table 4.3 shows the most popular users using these three measures. Users in italics appear on more than one top ten list, and users with ^{p} after their names are Portuguese-speaking. The most followed top ten consists solely of celebrities and news sources, and the top ten mentioned users are also mostly celebrities[3]. However, eight of the top ten retweeted users are faceless organizations which focus on propagating a specific meme or website. This shows that, although users follow and talk about celebrities, they don't find their

[2]At least on the first level of the retweet tree, but as these trees are very shallow, this level matters the most

[3]One of these two exceptions, `@dealsplus`, promises entry to a daily prize draw to anyone who tweets a specific message they provide, promoting their website and mentioning them, which obviously inflates their mention count

Rank	Follower Count	Retweets	Mentions
1	aplusk	ZodiacFacts	*justinbieber*
2	nytimes	iDoit2	ivetesangalo^p
3	eonline	RevRunWisdom	addthis
4	tonyhawk	*dealsplus*	*dealsplus*
5	PerezHilton	ihatequotes	joejonas
6	*justinbieber*	*justinbieber*	nickjonas
7	huckluciano^p	OhJustLikeMe	luansantanaevc^p
8	NBA	TheLoveStories	DonnieWahlberg
9	johnlegend	Sexstrology	pelurestart^p
10	brookeburke	VouConfessarQue^p	ladygaga

Table 4.3: Top 10 users using three influence measures

tweets as worthy of being spread as those of these single-purpose information generators and aggregators. This can be expected, in a way, as these users exists solely to focus on a specific topic and create interesting retweetable content.

As an interesting aside, the most mentioned celebrities tend to be significantly younger — the average age of top ten followed users is 34, compared to 24 for the most mentioned users. This is probably because the most active users on Twitter are younger than the general Twitter population, and they are interested (and therefore talk about) younger celebrities.

Chapter 5

Information Spread

This chapter moves beyond looking at individual tweets and users and examines how tweets and users can be related by retweet trees and graphs respectively, and the importance of topics. Language and topic divisions are examined in depth, and, using the metrics defined in Chapter 3, the differences across these divisions are shown and discussed.

5.1 Characterizing Information Spread

5.1.1 Retweet Trees

Retweet trees deal with the propagation of a specific tweet from user to user, and so are the most direct way to look at information spread. The root of each of these trees represents an original tweet, and all subsequent retweets are represented as nodes in the tree with edges going back to their parents. Most twitter retweet trees tend to be quite shallow, with most spread occurring from influential users tweeting or retweeting and then having this tweet retweeted by many of his or her followers. Figure 5.1 shows a typical retweet tree; most retweet trees follow this pattern — a shallow graph, where most retweets are in response to a few very influential users, which results in a characteristic star-like shape. This implies a lack of information cascades and therefore means that influential nodes are vitally important in spreading information. If we examine all the retweet trees within a given topic, the size distribution for small trees follows a power law, but more large trees than expected occur — these are probably due to the merging of smaller trees. A distribution of retweet tree size for a typical topic is shown in Figure 5.2.

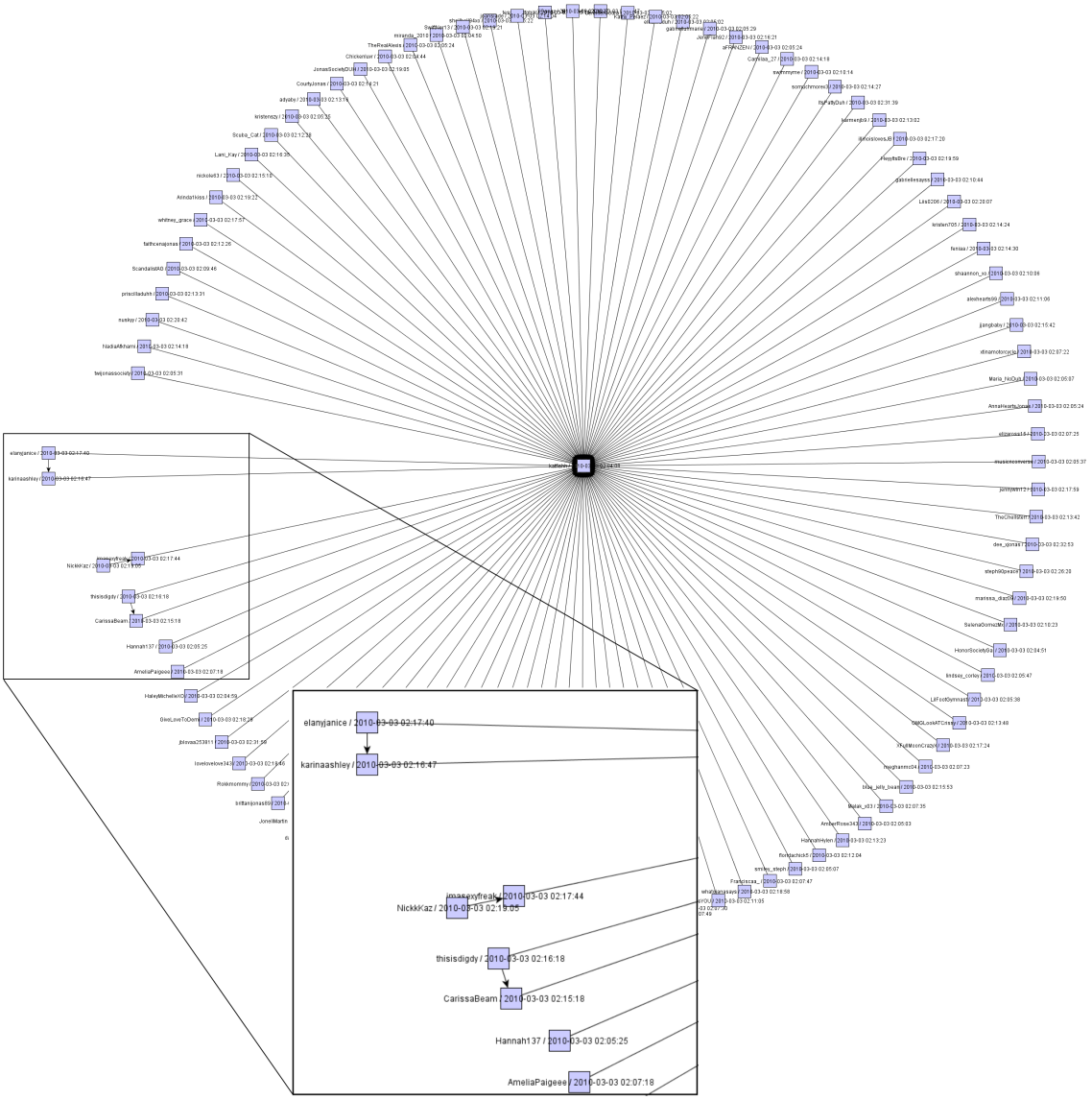


Figure 5.1: Example of a retweet tree (from trending topic `#RIPAlejandraJonas`)

5.1.2 Retweet Graphs

I use the term 'retweet graph' to refer to what is conceptually a sum and reduction of all the retweet trees pertaining to a single topic. Instead of nodes representing tweets, they represents users, and have in- and out-edges corresponding to the edges of all the tweets and retweets of that user. This leads to a directed graph, which is not necessarily or usually acyclic, representing all the retweet interactions in a specific topic over a period of time. Figure 5.3 shows an example of a retweet graph, taken from a set of tweets referring to the UK general election — it is typical of retweet graphs, and very different from a typical retweet tree. It is larger — and as the topic becomes broader or the timespan longer, it becomes even larger, eventually approximating Twitter's Giant Connected Component. It contains bidirectional links and long chains, showing that even though an individual tweet may not usually be retransmitted through many levels, ideas might be. The graph

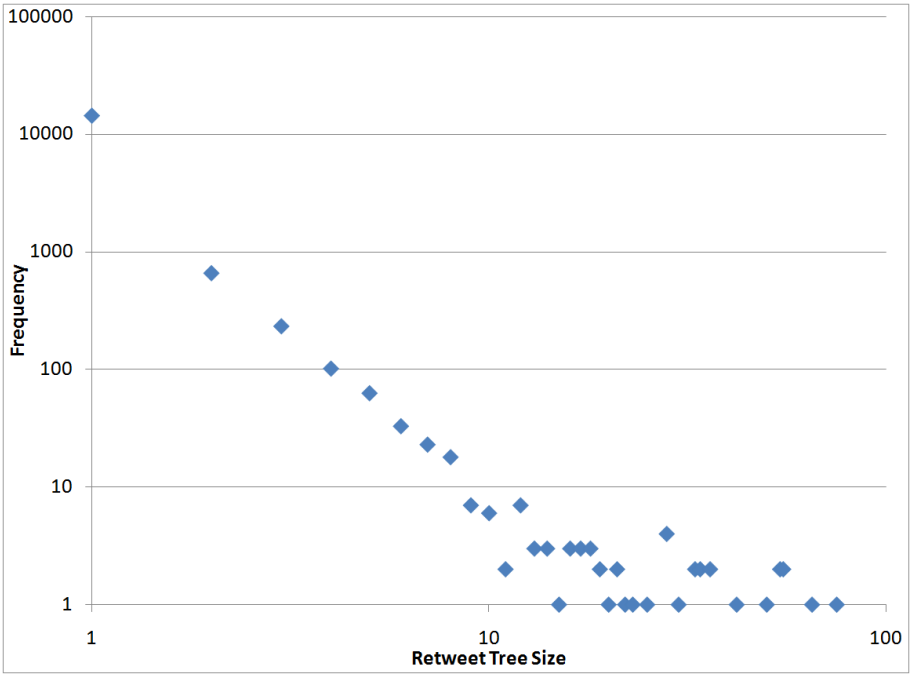


Figure 5.2: Retweet tree size/frequency for retweet trees in the topic 'Haiti'

is similar to retweet trees in that influential users have a very large presence in the graph; in this case, two users are, taken together, directly connected to most nodes in the graph.

5.1.3 Topics

Topics can be any word that commonly occurs in a tweet, and are not necessarily explicitly defined or invoked by users — although once a topic becomes popular, users may began to use it more consciously. Hashtags are used by convention to explicitly indicate or define topics, but non-hashtag topics are commonly amongst the most popular. As mentioned above, Twitter publicizes the top ten topics by region and globally through a constantly-updated list on the Twitter home page; topics which reach this level of popularity usually experience a further jump in popularity, in some cases accompanied by a geographical spread beyond initially limited audience. This effect also tends to help trending topics stay popular and leads to multiple popular hashtags collapsing into the most popular one.

5.2 Language

I chose to look at differences between speakers of different languages because language is the strongest barrier to intercommunication in Twitter, as its emphasis on global conversations allows participation in a topic to anyone who can understand the language

being used in that topic. Therefore I believed that dividing the sample tweets by language would lead to differences greater than, or at least comparable to, any other way of dividing the data. Language also provides a good degree of geographic locality — and, unlike location data which relies on often unreliable or missing user-provided metadata, it is easy to extract from the content of a tweet.

5.2.1 Language Popularity

Table D.13 shows my results, compared with existing work from Semiocast([Sem10]) and the WebEcology Project ([Bei10]).

Language	My Results	Semiocast	Web Ecology(Google)
English	55%	50%	62%
Japanese	14%	14%	6%
Portuguese	11%	9%	10%
Malay	10%	6%	3%
Spanish	5%	4%	3%
Other/Unknown	5%	17%	8%

Table 5.1: Comparison of language distribution results

5.2.2 Characterizing Languages

I examined four metrics — mentions, retweets, hashtags and URLs — in different languages: There are two ways to look at each of these metrics: as a frequency (total mentions/total tweets, or average mentions/tweet) or as a rate (number of tweets mentioning at least one user/total number of tweets). For the metrics for which there is a significant differences between rate and frequency I discuss both, but generally, and how they varied for different languages, rate is the more representative measure.

To help show the differences between these measures for different languages, in the graphs in this chapter the top language for each metric is colored, as well as English, the most popular language.

5.2.3 Individual Measures

Mentions

A higher level of mentions suggests more social or conversational use of Twitter, while fewer mentions suggest unidirectional content propagation with little 'backchatter' — so the higher the mention rate, the more users of are treating Twitter as a social network,

and the lower the rate, the more they regard it as an information source. The mentions rates show a very clear difference between languages, even within superficially similar geographic/cultural groups, such as Korean and Japanese — Korean tweets mention nearly twice as many users.

Retweets

Retweet rates show a higher inter-language variance. Malay, Chinese and Thai have a much higher retweet frequency than any of the other languages. A high retweet rate can correlate with social activity, as the main function of retweets is to spread information to followers; however, retweeting can also be used just to boost topic rankings or express agreement, both of which do not depend on social network links, although expressing agreement is a social interaction.

Hashtags

Hashtags show use of twitter for topical discussion, as opposed to communication with a slow-changing social network. Again, the average hashtag incidence differs significantly between languages. Malay comes last in this measure, whereas it came top in mentions per tweet. A possible explanation for a low level of hashtag use is that Twitter is used as more of a social network than a broadcast medium, and thus users tend not to use global hashtags as much as in regions where global conversations are the norm. However, this difference could just be due to the differences in the uptake of hashtags; as the function of a hashtag — defining a topic and possibly appearing on the trending topics list — can be performed by any phrase, hashtags may just have not become as popular in some of these regions.

A high hashtag rate could be a sign of a high level of 'spam' tweets attempting to get noticed by including multiple popular hashtags. A large difference between hashtag frequency and rate, as seen in German and Italian, is an indicator of this kind of spam.

URLs

A large number of URLs in tweets is indicative of users using Twitter as an information source rather than a means for conversation. It points to less discussion and more one-way information dispersal taking place and a focus on out-of-band (either real-world or Internet-based but outside Twitter) news, memes or information. A high URL rate could also be a due to a lack of a large enough number of users to form a real community; this would lead to little retweet or mention activity and thus indirectly raise the URL rate, explaining why the less popular languages exhibit higher URL rates.

5.2.4 Combined Measures

To get a good sense of how different languages compare, it is useful to look at all four factors in one graph. Figure 5.8 shows the five most popular languages and their normalized rates for each metric on a radar chart.

5.2.5 Potential Explanations

How can these inter-language differences be explained? There are several plausible hypotheses:

- Cultural differences: People living in different cultures use the Internet, and by extension Twitter, in different ways and with different goals[CCM+02]. Language serves as a useful way to coarsely identify certain of these cultures. In short, Indonesians, compared to Western Europeans, may be more inclined to chat with friends online rather than posting a link to their latest introspective blog entry.

- Differences in how different language/regional groups currently use the Internet: This idea is similar to the one above, but looks for a more proximate cause than cultural differences; namely, the take-up and use of other social networks/news sources within a particular language group or region. In areas where there is a dominant social network already present (other than Twitter), we may expect less use of twitter as a way of chatting with friends.

- 140-character limit: Twitter enforces a maximum message length of 140 unicode characters; logographic languages, such as Chinese and Japanese, can represent significantly more information in a single character than alphabet-based languages. This effectively makes the amount of information that can be put in a single tweet depend on language. In addition, as user names are written in standard English alphabet (plus numbers and '_'), mentions and retweets carry a greater comparative cost, in terms of information, in logographic languages.

- Population penetration: Twitter has different degrees of popularity depending on language[1], and this disparity may affect how people use twitter: for example, many western european languages are relatively rare on twitter, and also tend to consist more of news and links rather than conversations; this could be a causal link.

- The interface with which users interact with Twitter: As shown in Figure 5.9, there are significant differences in what devices are used for Twitter across different languages. This affects these metrics: for example, it is much harder to share a URL

[1]Explaining these differences is beyond the scope of this project; it is probably due to differing completive landscapes in the social networking sphere in different regions and network effects

using SMS compared with a smartphone or a PC, and using programs on interfaces which provide easy retweet capabilities and access to Twitter search and trends makes retweeting and hashtag use easier and more appealing. This theory reverses the direction of causation; instead of people using Twitter because they use social networks of the Internet differently, people may develop habits and use patterns from their use of Twitter on less powerful devices that may then affect their future use of the Internet.

5.2.6 Future Directions

Although offered several ideas in the above section, to discover exactly why and how people use Twitter differently, a user study of users with different native languages would have to be performed. Examining how and why these users use Twitter, what other social networks and information sources they use, and what devices or services they use to access Twitter would provide insights into why the observed metrics differed between languages.

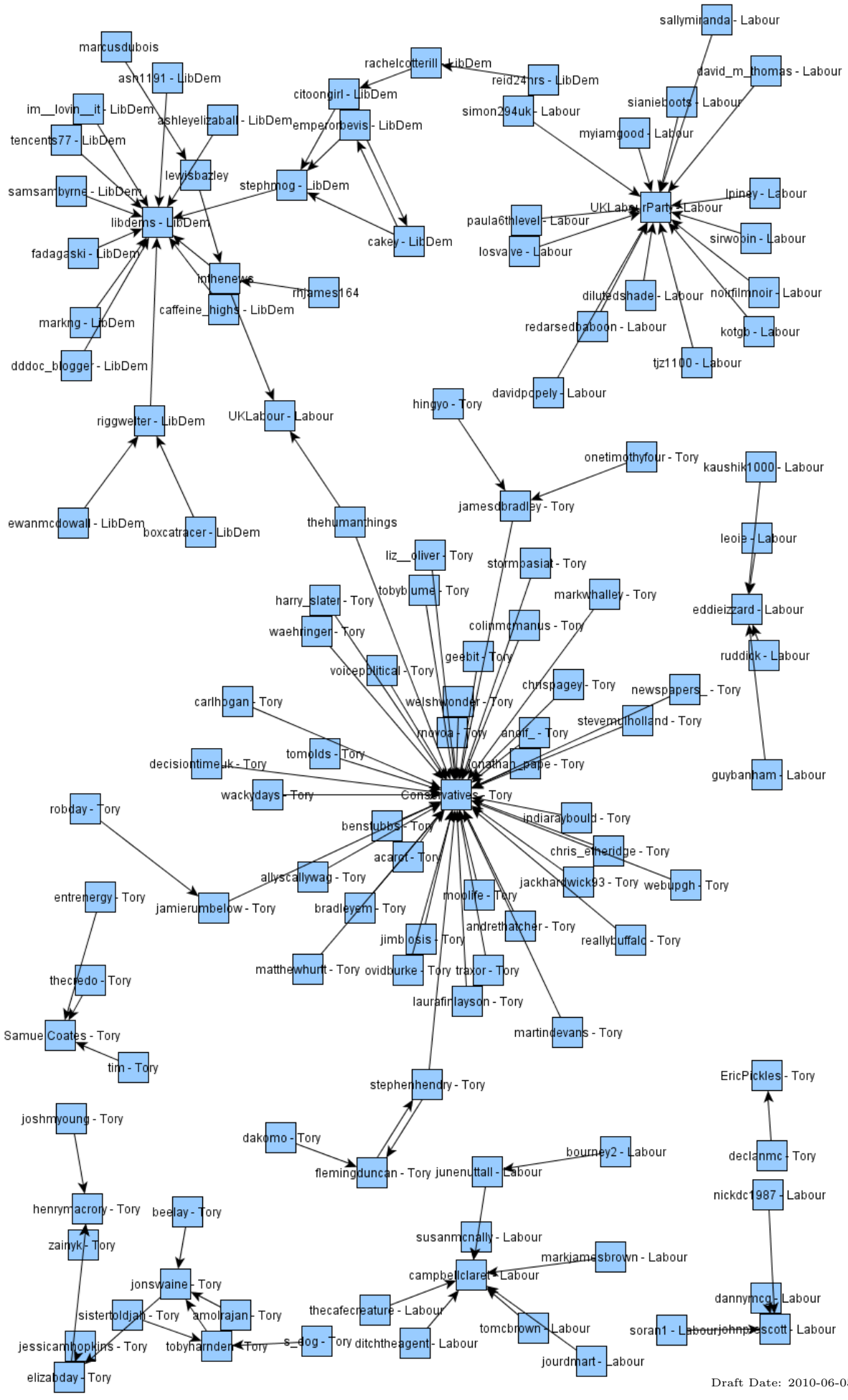
5.3 Topic

The same techniques as described in the previous section can also be used to investigate differences between conversations about different topics. Again, a higher proportion of mentions shows a more conversational, social, two-way exchange of information, and is often seen for non-newsworthy generic topics such as `#GoodMorning`. A higher proportion of retweets show fast, directed spread of information, and are characteristic of fast-moving news events, such as `Pacific Tsunami`. High use of hashtags is common in temporary, global conversations, such as `Olympics`. A large number of URLs is indicative of either spam (`ViagPure`) or internet/technology news such as `iPad`. With the large number of categories used for this manual analysis, it's very difficult to accurately categorize topics. However, in limited domains good accuracy is possible; for example, looking just at URL and mention rates (Figure 5.11) we can see interesting results:

- Technology News: Technology related news stories show consistently high URL frequencies; this is not surprising, as technology news stories tend to be more web-based than other news.
- Movies/TV Shows: Movies as opposed to TV shows showed higher URL rates: this could be because of movies tend to be more temporal and newsworthy topics while TV shows are usually a topic. In Figure 5.11 there is a single outlying 'Movie' data point with a very high URL rate; this is the topic `Tron Legacy`, which has a strong following amongst the tech community; this somewhat correlates with the above observation that technology related news has a higher URL rate.
- Memes: The 'Meme' category is a catch-all for topics which were not related to any non-Twitter event or anyone in the real world — this includes content-poor topics like `GoodNight` and original Twitter-based content like `#UKnoBLAH`. These topics understandably showed the lowest URL frequencies, and a wide range of generally high mention rates.

5.3.1 Topic Categorization

Being able to automatically categorize topics has implications for recommendation engines(such as the one in [CNN+10]), content archiving, and intelligent aggregation/presentation of popular topics. As seen in Figure 5.11, even looking at only two variables we can begin to separate certain similar categories — if all the four metrics, as well as tweet length, language and the source of the tweets (mobile, smartphone or PC) are taken into account, a quite accurate categorization system might be feasible.


sallymiranda - Labour
marcusdubois
rachelcotterill - LibDem
reid24hrs - LibDem
david_m_thomas - Labour
ash1191 - LibDem
citoongirl - LibDem
sianieboots - Labour
im__lovin__it - LibDem
ashleyelizaball - LibDem
emperorbevis - LibDem
simon294uk - Labour
myiamgood - Labour
tencents77 - LibDem
lewisbazley
samsambyrne - LibDem
stephmog - LibDem
lpiney - Labour
UKLabourParty - Labour
libdems - LibDem
paula6thlevel - Labour
sirwobin - Labour
cakey - LibDem
losvave - Labour
fadagaski - LibDem
inthenews
rhjames164
dilutedshade - Labour
noirfilmnoir - Labour
caffeine_highs - LibDem
markmg - LibDem
redarsedbaboon - Labour
kotgb - Labour
dddoc_blogger - LibDem
tjz1100 - Labour
davidpopely - Labour
hingyo - Tory
riggwelter - LibDem
UKLabour - Labour
onetimothyfour - Tory
kaushik1000 - Labour
jamesdbradley - Tory
ewanmcdowall - LibDem
boxcatracer - LibDem
thehumanthings
leoie - Labour
liz__oliver - Tory
stormbasiat - Tory
tobybaume - Tory
markwhalley - Tory
harry_slater - Tory
eddieizzard - Labour
colinmcmanus - Tory
waehringer - Tory
ruddick - Labour
geebit - Tory
voicepolitical - Tory
chrispagey - Tory
newspapers_ - Tory
carlhogan - Tory
welshwonder - Tory
stevemcilholland - Tory
movoa - Tory
anoif_ - Tory
tomolds - Tory
decisiontimeuk - Tory
jonathan_pape - Tory
guybanham - Labour
wackydays - Tory
Conservatives - Tory
robday - Tory
indiaraybould - Tory
benstubbs - Tory
chris_etheridge - Tory
acarrot - Tory
allyscallywag - Tory
jackhardwick93 - Tory
webupgh - Tory
entremergy - Tory
moolife - Tory
jamierumbelow - Tory
bradleyem - Tory
andrethatcher - Tory
jimbosis - Tory
reallybuffalo - Tory
matthewhurst - Tory
ovidburke - Tory
traxor - Tory
laurafinlayson - Tory
thecredo - Tory
martindevans - Tory
Samue Coates - Tory
tim - Tory
EricPickles - Tory
stephenhendry - Tory
joshmyoung - Tory
dakomo - Tory
declanmc - Tory
bourney2 - Labour
flemingduncan - Tory
junenuttall - Labour
nickdc1987 - Labour
henrymacrory - Tory
beelay - Tory
susanmcnally - Labour
zainyk - Tory
markjamesbrown - Labour
jonswaine - Tory
campbellclaret - Labour
dannymcg - Labour
sistertoldjah - Tory
amolrajan - Tory
thecafecreature - Labour
soran1 - Labour
johnprescott - Labour
jessicamhopkins - Tory
tobyharnden - Tory
s_dog - Tory
ditchtheagent - Labour
tomcbrown - Labour
elizabday - Tory
jourdmart - Labour

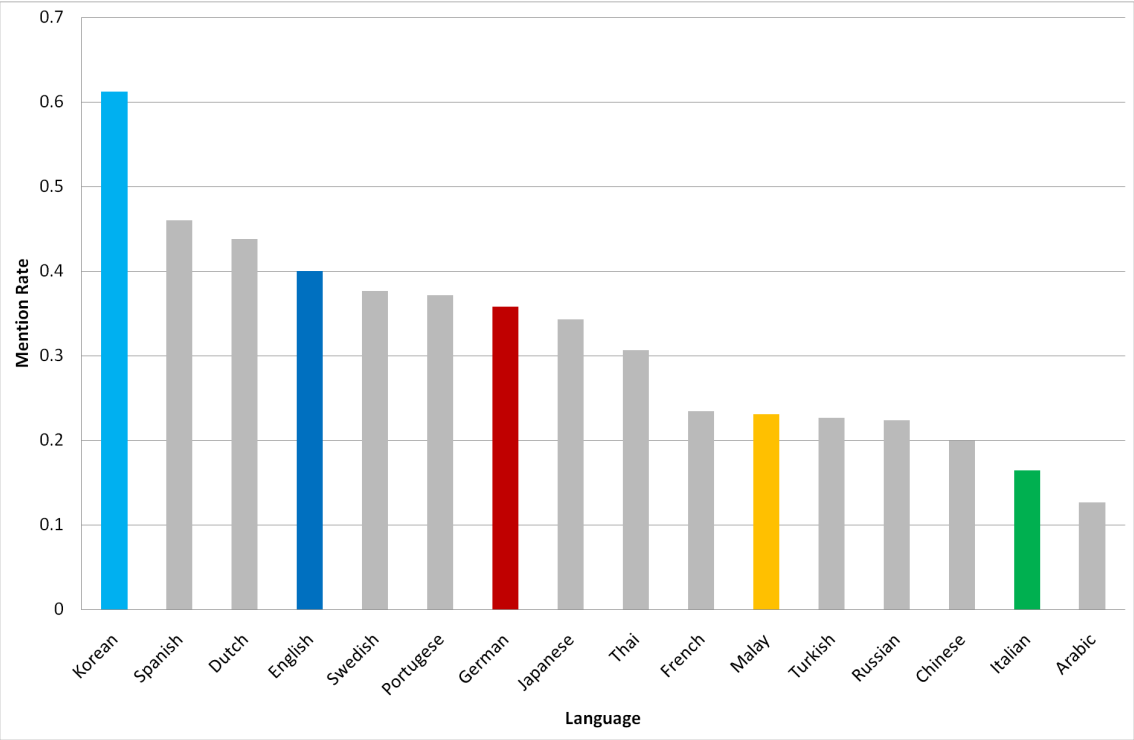


Figure 5.4: Mention rate and frequency for the 12 most popular languages

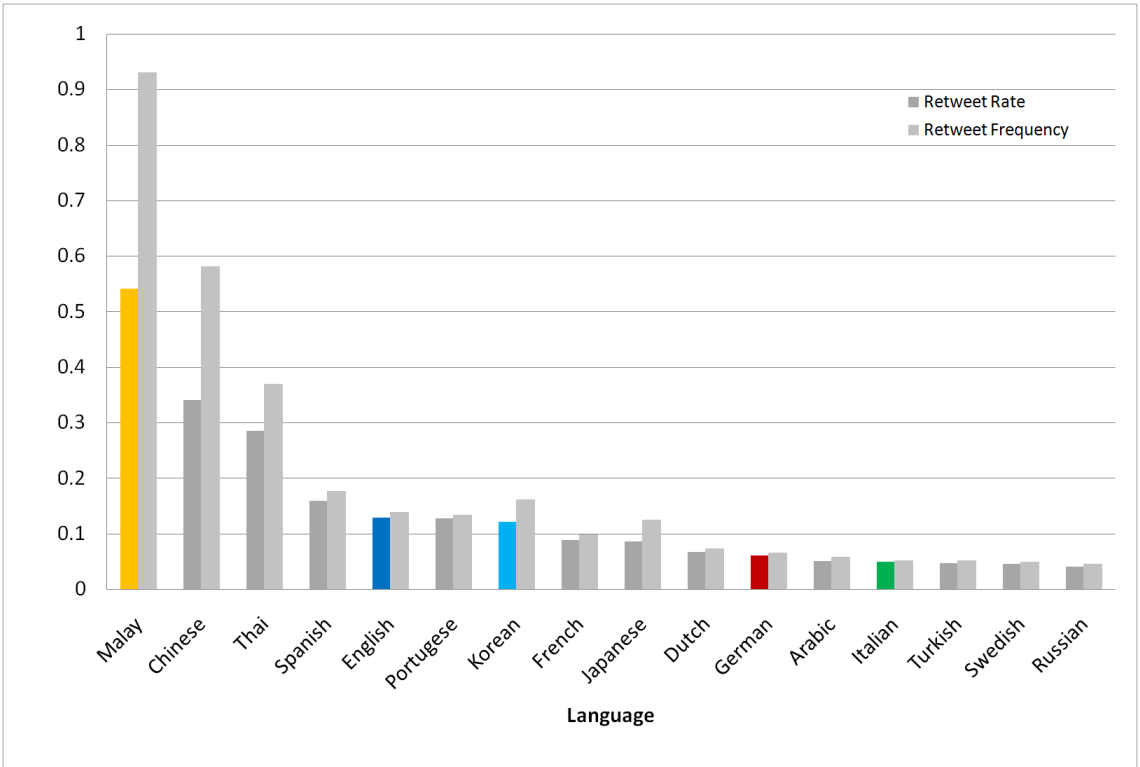


Figure 5.5: Retweet rate for the 12 most popular languages

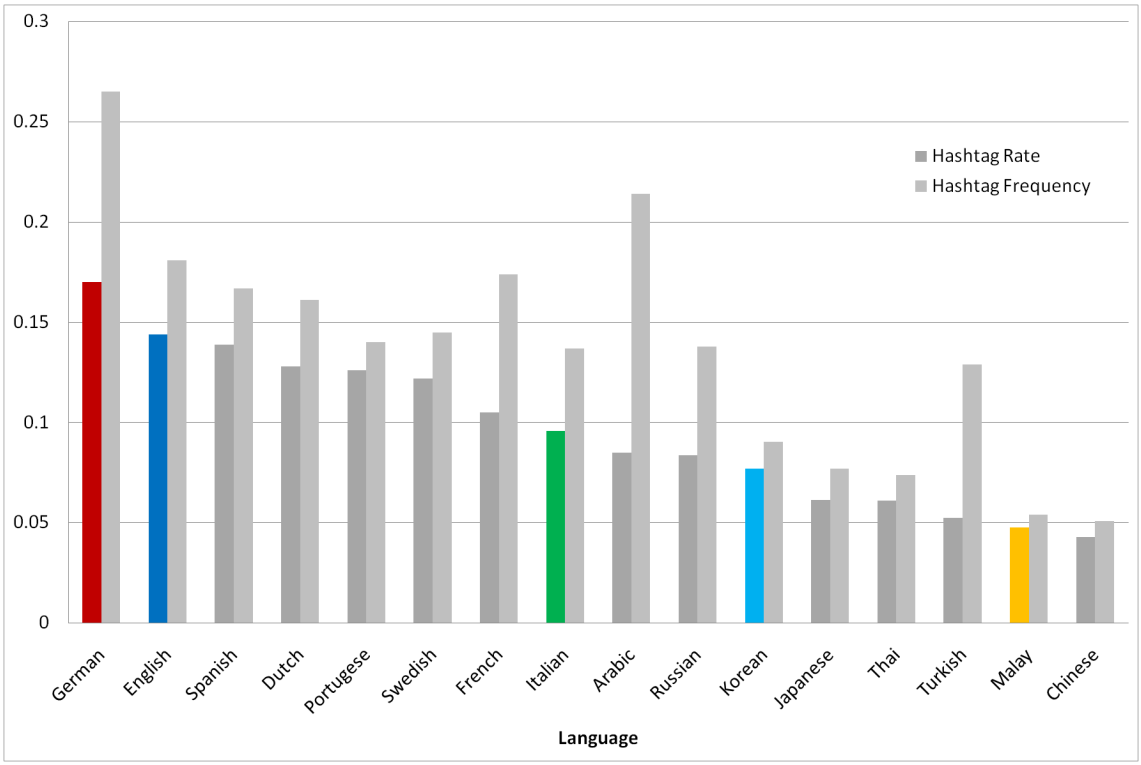


Figure 5.6: Hashtag rate and frequency for the 12 most popular languages

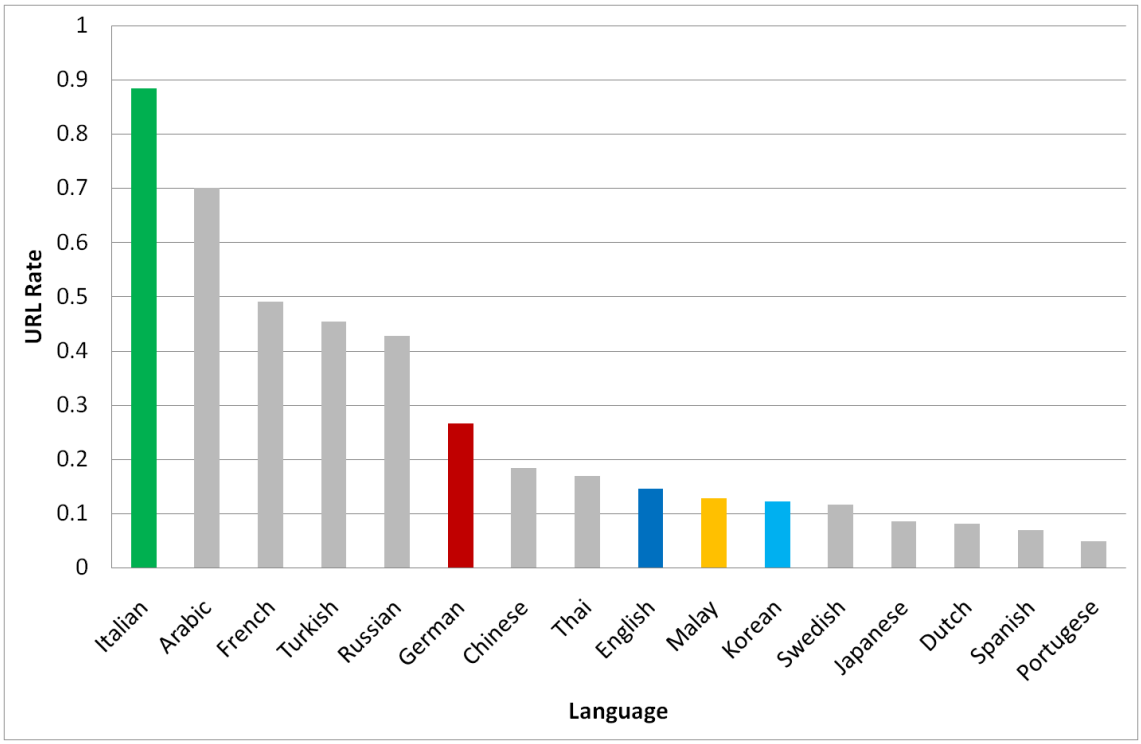


Figure 5.7: URL rate for the 12 most popular languages

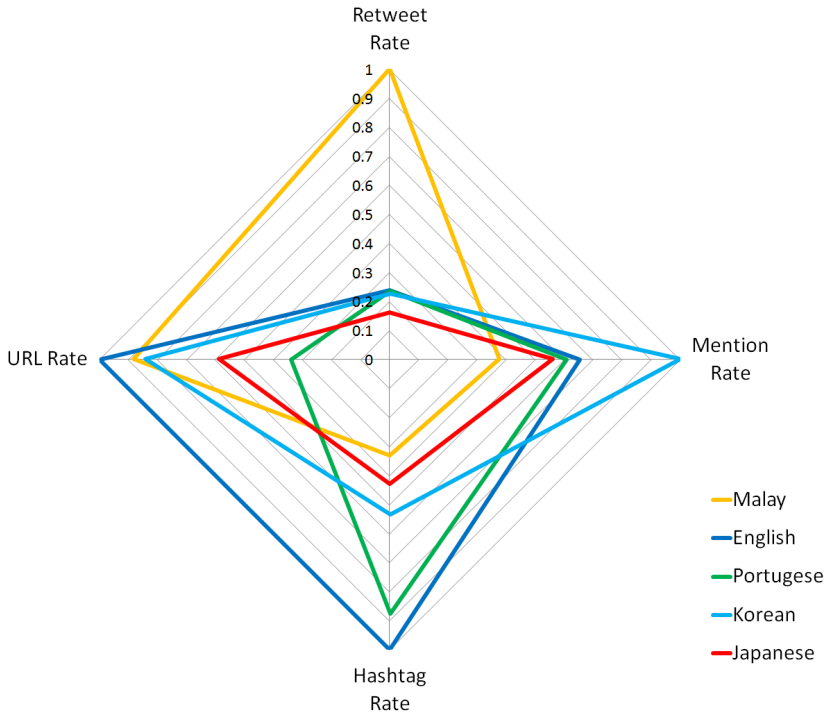


Figure 5.8: Combined (normalized) metrics for the five most popular languages

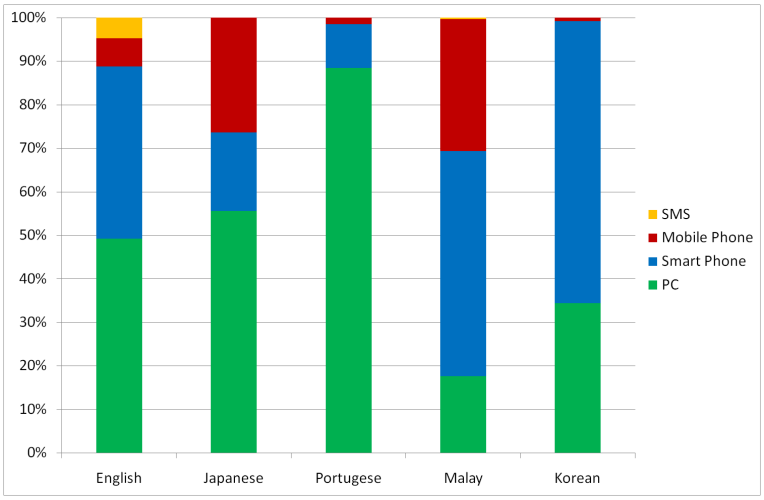


Figure 5.9: Tweet source distribution for the four most popular languages

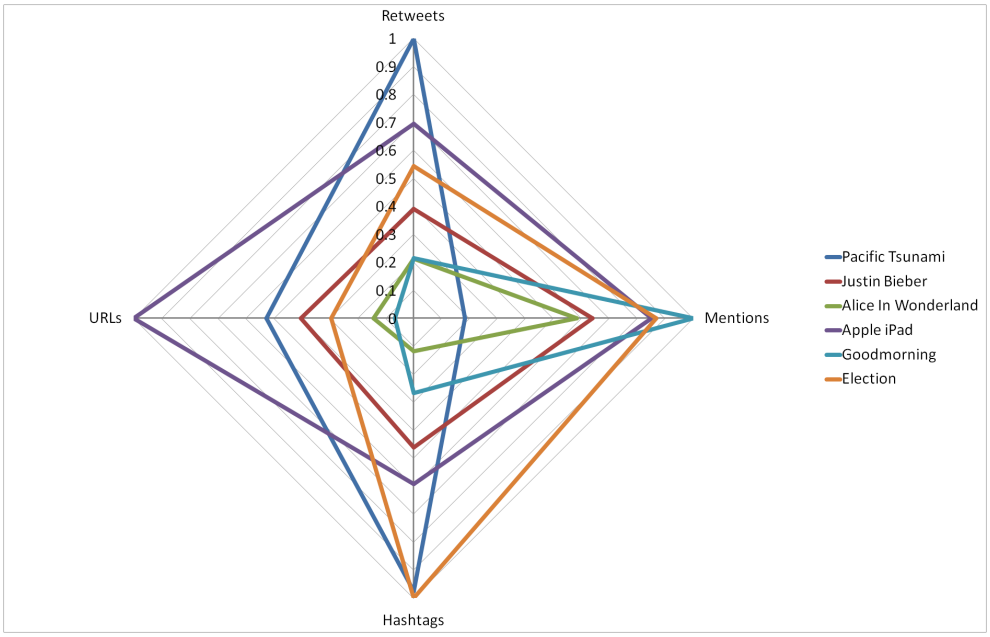


Figure 5.10: Differences between five popular topics

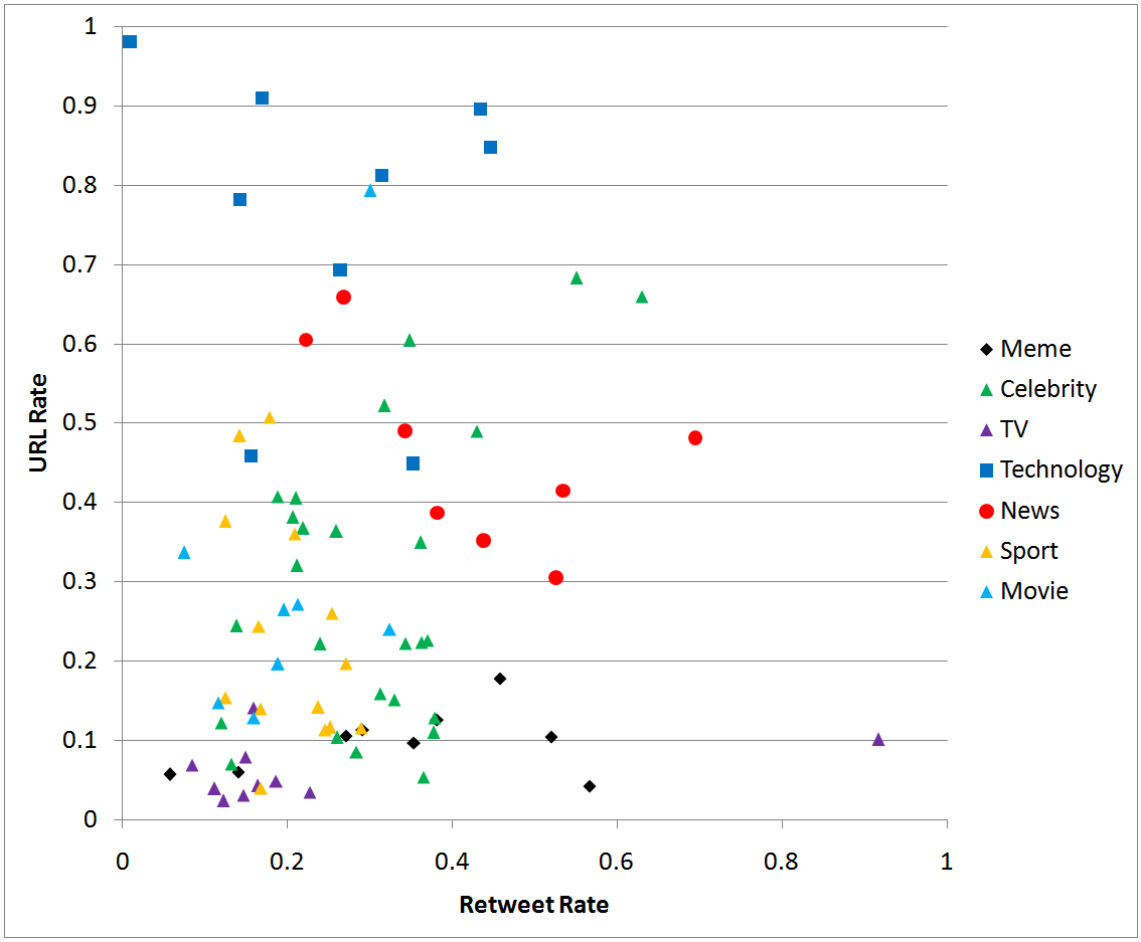


Figure 5.11: Scatter graph showing manually categorized topics

Chapter 6

Case Study: UK General Election

This chapter presents a case study of the tweets related to a single set of topics related to a major real-world event: the 2010 UK general election. By looking at temporal changes in activity, Twitter is shown to react instantly to real-world events. Basic textual and semantic analysis shows how the general views of the user population can be extracted and examined, and a method for deducing users' party affiliation is presented.

While the election was used because of the high volume of data relating to it, these methods could be used to examine brand ties and allegiances, monitor the different associations formed by users between different products, and examine how real-world events such as product launches affect the volume of related Twitter activity.

6.1 Timeline

The UK General Election took place on May 6th, 2010, and contested by three major parties: the Labour party, the incumbents, led by Gordon Brown, the Conservative Party, led by David Cameron, and the Liberal Democrat party, led by Nick Clegg. Although exit polls and initial results were released on the night of the 6th, the final outcome of the election, due to the UK parliamentary system, was not clear until the 11th of May, when Gordon Brown resigned and David Cameron become prime minister, announcing that he would attempt to form a coalition with the Liberal Democrats. Major events that occurred over the period I was recording include:

- *22:00, May 6th* — Exit polls released indicating a hung parliament.
- *22:55, May 6th* — Labour holds the first seat to report, although there is a large swing in voter share to the Conservative party. The next two seats report at 23:30 and 23:41, and also are Labour holds.
- *01:05, May 7th* — The Conservative party wins its first seat.

- *01:16, May 7th* — The Liberal Democrat party wins its first seat.
- *01:38, May 7th* — Gordon Brown holds his seat and gives a speech.
- *03:05, May 7th* — David Cameron holds his seat and gives a speech.
- *06:50, May 7th* — Nick Clegg holds his seat and gives a speech.
- *10:50, May 7th* — Clegg says that Conservative party deserve the first opportunity to form a government.
- *13:45, May 7th* — Gordon Brown speaks, raises possibility of talks with Liberal Democrats.
- *14:40, May 7th* — David Cameron speaks and publicly offers deal to the Liberal Democrats.
- *15:52, May 8th* — Nick Clegg addresses a crowd of demonstrators in London
- *17:00, May 10th* — Gordon Brown gives a speech, stating his intention to resign within a year, announcing formal talks with the Liberal Democrats, and suggesting a "progressive" coalition.
- *18:20, May 10th* — Nick Clegg appears on television and welcomes Brown's announcements.
- *19:20, May 11th* — Gordon Brown resigns.
- *20:26, May 11th* — David Cameron becomes Prime Minister.
- *20:45, May 11th* — Cameron arrives at 10 Downing Street and gives a speech.

6.2 Temporal Analysis

Looking at the party-related activity over the election shows that Twitter users react very quickly to newsworthy events. In Figure 6.1[1], which shows the entire week around the election, major events can easily be picked out; for example, the peaks above 1000 tweets per ten minutes, from left to right, correspond with the exit polls being released (hung parliament predicted, Liberal Democrat spike) and Labour winning the first three seats (three closely-spaced Labour peaks), both on election night, Cameron giving a speech offering to work with the Liberal Democrats the next day, and finally Gordon Brown's speech on May 10th suggesting a progressive coalition and stating he would step down within a year (Labour/Liberal Democrat spike). Surprisingly, none of the events of May

[1] A larger version of these graphs is in Appendix C

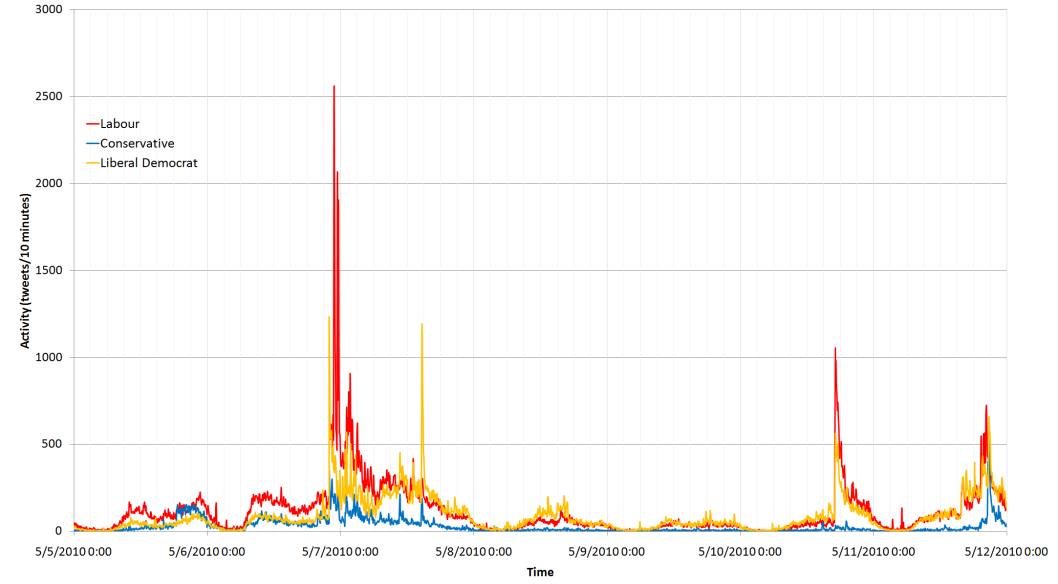


Figure 6.1: Party mentions during the week around the election

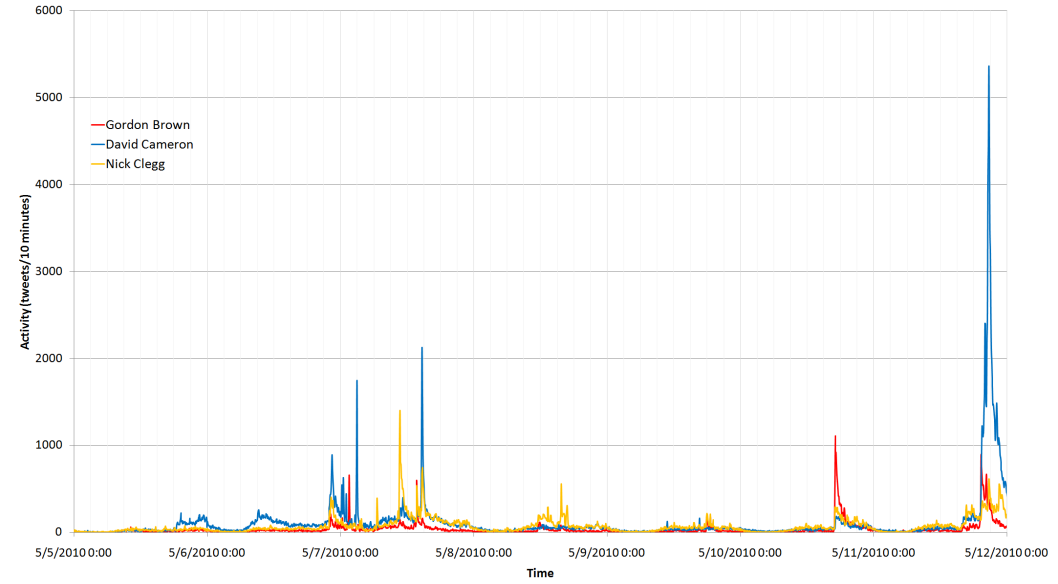


Figure 6.2: Party leader mentions during the week around the election

11th — Gordon Brown's resignation, the coalition deal and David Cameron becoming Prime Minister — reach the level of activity. This is due to increasing focus on 10 Downing Street in the aftermath of the election; the party supporters had stopped celebrating or eulogizing, and the focus was on the personalities, not the parties. This is supported by Figure 6.2, which shows a massive (5363 tweets in a ten-minute period) and long (over 100 tweets/minute for over 3 hours) peak in mentions of David Cameron. An interesting general effect is that while the Labour party was the subject of far more tweets than the other two parties, both David Cameron and Nick Clegg showed more activity than Gordon Brown: this may be because Labour supporters are more likely to tweet about opposing candidates than opposing parties, a result of Gordon Brown being the incumbent, or due

to users, regardless of party, not being interested or excited about Brown.

Looking at Figure C.1, which shows party leader Twitter activity on the day after the election, along with the changes in the FTSE 100 index over the course of the day, it seems that the FTSE 100 lagged behind Twitter in responding to news events: It drops sharply after Gordon Brown's speech making clear his intentions to try and form a government and recovers after David Cameron offers a deal to the Liberal Democrats[2]. Both this graph and the ones in Figures 6.1 and 6.2 show that newsworthy events are not only immediately reported on Twitter, but reported and retweeted many times, leading to a detectable burst in activity. Being able to accurately detect these events in real-time could have uses in automatic news tracking/notifications, or even, as hinted at above, automated trading platforms.

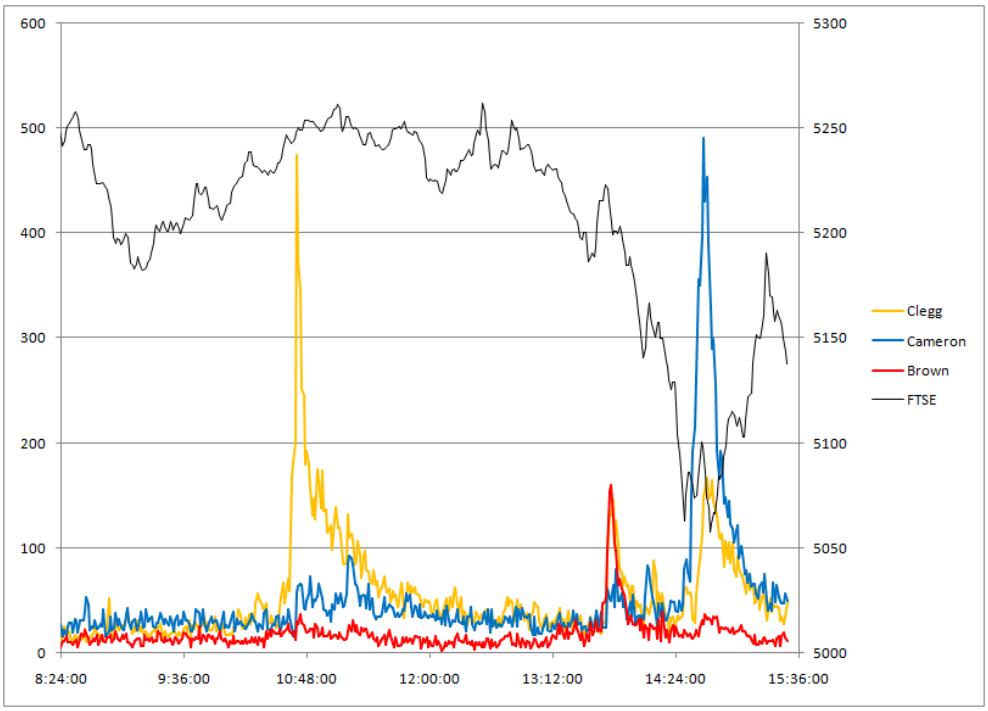


Figure 6.3: Party leader mentions and the FTSE 100 Index during the day after around the election

6.3 Party Affiliation

Figure 6.4 shows a portion of the retweet graph within election-related tweets from a 5-day period after the election. The graph is rooted on the party affiliation-seeded users (given in Appendix B.1), so only users from which these seeds are reachable through any number of retweet links — not necessarily of the same tweet — appear on the graph. Despite this, 17,835 nodes appear on in the graph, though only 1,422 are colored — showing that

[2]Of course, this is only one possible interpretation of market movements

most retweeters are not pure partisans, and retweet users that don't necessarily follow the same party line. By identifying nodes in this graph with a high in-degree, influential users and the parties they support can be found. On a similar graph (shown in Figure C.5 for which the affiliation propagation was limited to just two jumps from the seed nodes, 24% of the nodes are connected to at least one of the seeds, and of these 82% are connected to seed nodes from all three parties, showing that, even with the two-hop limitation, most retweeting users cite information from partisans of all sides. Table 6.1 shows the number of users in each category, for both graphs — interestingly, there are more users which were associated with both Labour and the Liberal Democrats than associated with the Conservatives and Liberal Democrats; this despite the fact that the Liberal Democrats formed a coalition with the Conservatives (This sample was taken after the election, and the numbers who supported the Liberal Democrats alone or the Conservatives alone are comparable). This is obviously a very simple technique, and is intended only as a proof-of-concept. I plan to extend it to take into account sentiment and frequency of users retweeting known party supporters to assign affiliation probabilities, instead of the current all-or-nothing approach, which can miss valid supporters if they happen to retweet

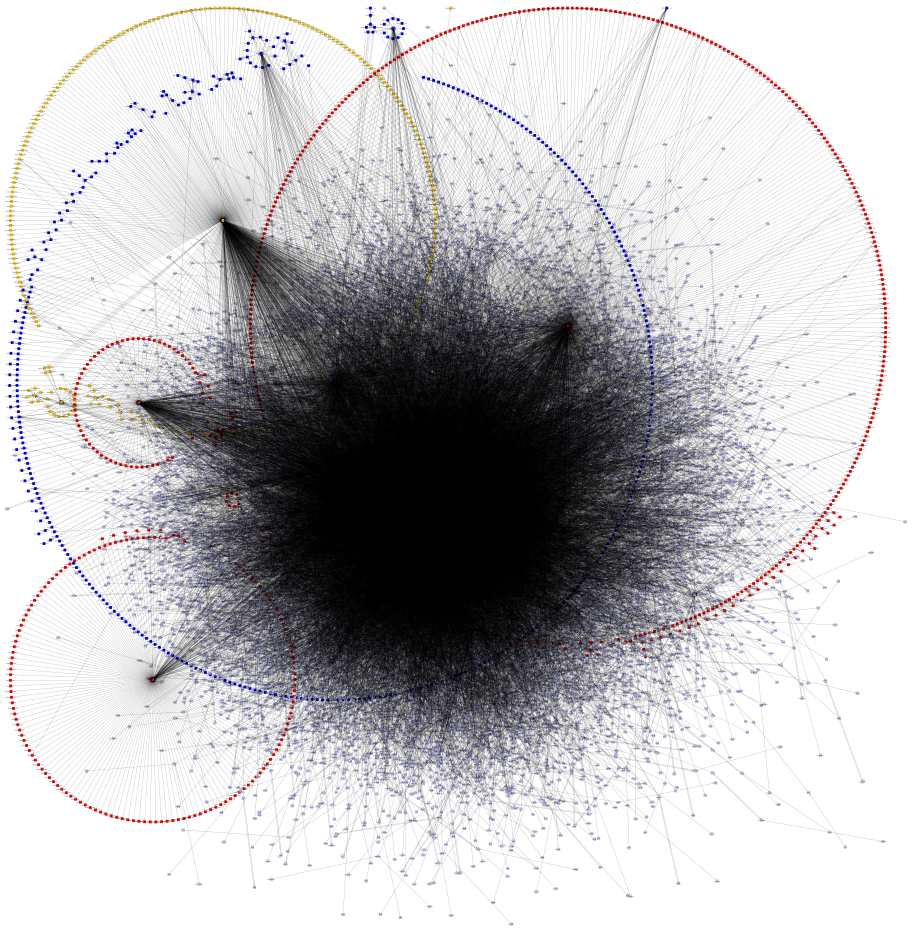


Figure 6.4: Graph colored by party affiliation from a 157,764 tweet sample from 05:43 7/5/2010 to 23:17 11/5/2010

Affiliation	Count (until stable)	Count (2 hops)	Percentage (until stable)	Percentage (2 hops)
None	51,899	53,184	74.4%	76.3%
All three	16,362	13,490	23.5%	19.3%
Labour	686	1654	1.0%	2.3%
Conservative	429	495	0.62%	0.71%
Liberal Democrat	307	427	0.44%	0.61%
Labour and Liberal Democrat	20	347	0.03%	0.50%
Labour and Conservative	10	85	0.01%	0.12%
Conservative and Liberal Democrat	21	52	0.03%	0.07%

Table 6.1: Party affiliation assigned, based on seeds given in B.1, from a 157,764 tweet sample from 05:43 7/5/2010 to 23:17 11/5/201

a single tweet from a supporter of a different party, and can miscategorize objective users such as news sources[3] that retweet political figures. This technique could also be used with companies or products instead of political parties, allowing automated discovery of loyal supporters or customers or users who actively dislike companies or products. Combined with the influence measures described in Chapter 3, this could be used to very targeted marketing — for example, offering supporters the opportunity to test new products, or providing free products to detractors in an effort to 'convert' them.

6.4 Parties and Buzzwords

By examining relative rates at which keywords were used in conjunction with party or politician names, a snapshot of people's views and opinions about these parties or politicians can be obtained. Figure 6.5 shows the occurrence of various political buzzwords in conjunction with the three main party leaders. Understandably, 'change' is mentioned the least in conjunction with Gordon Brown, the incumbent. However, 'future' is mentioned the most in conjunction with Gordon Brown. This could indicate differences in the way that users talk about the next few years depending on their party affiliation — Labour supporters, instead of using the somewhat changed and anti-incumbent 'change', talk about the more vague concept of 'future'. Figure 6.6 shows how often the party leaders are mentioned in conjunction with the names of the two most recent US presidents in the days before the election. Nick Clegg's campaign was compared extensively to Obama's,

[3]For for the purpose of this example, I'm assuming that news sources are objective

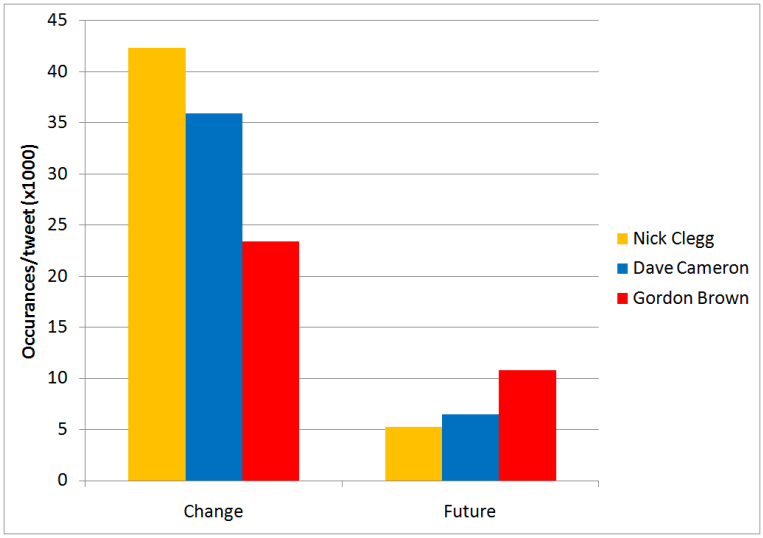


Figure 6.5: Pre-election coincidence of keywords and party leaders

as was David Cameron's, but the Clegg comparisons seem to have resonated more with Twitter users.

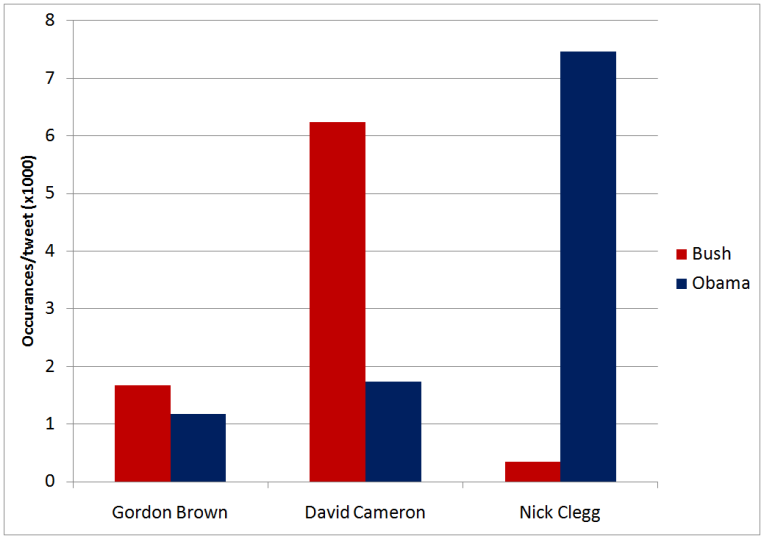


Figure 6.6: Pre-election coincidence of american politicians and party leaders

6.5 Sentiment Analysis

In the days before the election, I analyzed the activity levels for the various parties contesting the election. However, measuring raw activity is misleading, as some parties receive a good deal more criticism on Twitter than others — and, in politics, all publicity is not good publicity. By analyzing the context in which users mentioned voting for each party, a more accurate picture can emerge, accounting of course for the inherent biases present due to Twitter's userbase. Figure 6.7 shows the results: unsurprisingly, the BNP fares the

worst, as it had been receiving disproportionate attention relative to its support, most of it negative. The Conservative party follows, as expected from the overrepresentation of Labour support on Twitter. Interestingly, there are a lot of ambiguous tweets referring to the Liberal Democrats, possibly an indication of the uncertainty of the voter about about the effects of voting for a party other than the big two.

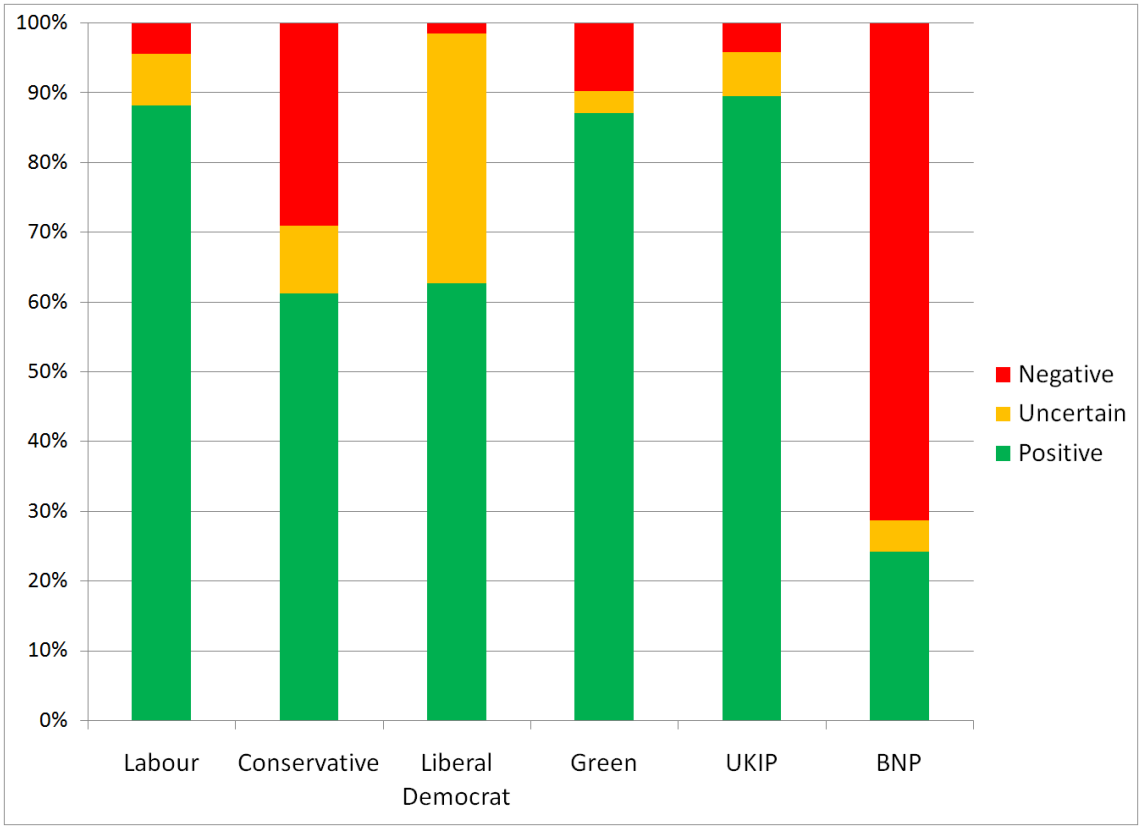


Figure 6.7: Twitter users positive/negative views towards the parties in the day before the election

6.6 Summary

The four methods presented in this case study can be applied to analyze other newsworthy events, or, over a longer timescale, analyze slower-moving trends. This latter application is of more use in a marketing context where, except for very large companies, relevant Twitter activity is fairly low. Correlating temporal analysis of topics related to a product or brand with real-world events can help evaluate marketing campaigns or model what effect unrelated, but major, real-world events have on user enthusiasm. Buzzwords and sentiment analysis can help companies discover what customers associate their products with, and how they feel about them and their competitors. Finally, affiliation propagation and analysis lets companies identify loyal users and target them more effectively.

Chapter 7

Conclusions & Future Directions

7.1 Conclusions

This project has shown that significant inter-language and inter-topic differences exist in four basic metrics used to characterize Twitter data. The work presented in chapter 5 explains the use of these metrics to characterize data, shows the magnitude of these inter-language differences and explores their implications, while chapters 3 and 4 present the techniques used to acquire and analyze the data, and show, and compare with other relevant research, general empirical results.

Chapter 6 shows some of the large variety of practical data which can be extracted from close examination of a single topic. The methods presented could prove useful in a commercial context, whether for identifying brand-loyal (or disloyal) users or finding out how users think of a product compared to the competition. This has clear applications in advertising design and targeting.

To answer the question implied in the introduction, and by [KLPM10], namely 'is Twitter, a social network or a broadcast medium?', is not straightforward. Right now, it is both, and how where it lies in between the two options depends on language and topic. However, looking at the changes in Twitter over the past year, both in the way users access it and the changes to the website emphasizing search and trends, Twitter is positioning itself as an information source and de-emphasizing its social aspects. This change will make most conversations more global, and make advertising easier, as users will not care as much about the source of tweets, just their content — Twitter seems to recognize this, and has introduced sponsored tweets on its search pages, something that would not be feasible a few years ago, when search did not exist and most users were interacting with Twitter via SMS. However, this change will not necessarily be fast, and may not even take place, everywhere. Regions, isolated by language from the rest of Twitter, may well continue to place their emphasis on social links, especially in places where mobile devices are still the main way users are accessing Twitter.

Treating Twitter as if it is a single large homogeneous community is misleading; there are many possible factors, such as location, age, education, or language which could affect the behavior of users and divide them, to a lesser or greater extent, into separate communities. From these, I focussed on language, because language divisions breaks up the Twitter user-space cleanly into mostly disjoint sections, connected only by a small proportion of multilingual users and the global trending topics list. In practice, each of these language blocks regards Twitter in a different way, as a social network, an information source, or something in between. The different means by which users can browse, search and tweet also affects their use of the service, and varies greatly by region. All this must be taken into account when designing strategies for communicating with users in varied regions effectively.

7.2 Future Directions

7.2.1 Using Location Information

The same type of analysis that was carried out for languages and topics would naturally extend itself to location-based grouping. This might provide greater differentiation than broad language-based groups, although — due to the geographically wide-ranging social connections of most users — I doubt it will provide clear differences except in clear-cut distinctions such as UK-US or Portugal-Brazil. There are also difficulties in automatically detecting location, as many users do not provide their location, or provide a fake or useless location such as "The World"[1] or "justin bieber land". Twitter has begun to provide a per-tweet geotagging option (as opposed to the previous user-granularity, user-specified location) which may prove very useful in future; however, currently significantly less than 1% of tweets include geolocation information[2].

In combination with language data, robust location information would be very useful in identifying potential "bridge" users between different regions or languages. Users who tweet from multiple distant locations, or who tweet in languages which are not dominant in their region are more likely to spread language or region specific topics into new areas.

7.2.2 Multicasting, Channeling, and Aggregating

Twitter provides two methods for multicasting: hashtags, which associate the tweet with a global topic with fluid and changing population of users viewing it, and following, which allows users to statically opt-in to seeing a particular user's tweets. These are usually used

[1] Which, using Google's geolocation service, resolves to a building in New York City

[2] In addition, a non-negligible part of this use is non-standard: for example, automated earthquake notification systems which geolocate earthquake information with the epicenter

in conjunction, and users that contribute useful information within a topic are likely to be followed by people searching or tracking that topic, leading to the temporary link between the users becoming a more permanent and explicit follower-follow connection. Users that post often about different topics are likely therefore to have followers or friends associated with each one, and who care more about the topic that led them to the user in question. It follows that users are more likely to retweet tweets about the topic they are interested in, and this leads to implicit 'channels' within the Twitter social graph. These channels can be thought of as a subset of the graph for each topic or subtopic: a simple example is a clique of friends, one of whom is an expert on a topic; the implicit channel graph in this case is just the subset of the clique consisting of directed edges from the expert to her friends.

By analyzing the retweet activity of followers over we can possibly deduce how they 'met' the user they are retweeting, and which topics they retweet from that user. Users who have acquired followers from a certain topic, and have then interested these followers in unrelated topics (evidenced by them retweeting his tweets on that topic) are probably rare and influential, and worth investigating.

In a broader sense, the implications of this selective retweeting by followers is interesting, as it acts as a filter: for example, if `@alice` is currently following the BBC news twitter feed, but only cares about, and retweets, sports news, `@bob`, who is only interested in important sports news, can just follow `@alice`, who is not really acting as a content provider — as she's not providing anything not already available on Twitter — but as a filter([ZR09]). It is easy to extend this example further to more specific categories. As well as this, many users act as aggregators on a specific topic. Multiple levels of aggregation and filtering combined can lead to a quite comprehensive but concise feed of tweets focussed on a particular topic, and could be part of the appeal of Twitter.

Bibliography

[Bei10] Jon Beilin. Language detection and translation. `http://tinyurl.com/lehoe5`, 2010.

[BGL10] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS-43*, Kauai, HI, USA, January 2010. IEEE.

[BRCA09] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In *IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 49–62, New York, NY, USA, 2009. ACM.

[CCM+02] Patrick Y. K. Chau, Melissa Cole, Anne P. Massey, Mitzi Montoya-Weiss, and Robert M. O'Keefe. Cultural differences in the online behavior of consumers. *Commun. ACM*, 45(10):138–143, 2002.

[CHBG10] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington DC, USA, May 2010.

[CMG09] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 721–730, New York, NY, USA, 2009. ACM.

[CNN+10] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 1185–1194, New York, NY, USA, 2010. ACM.

[CT94] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[Dal09] Elizabeth M. Daly. Harnessing wisdom of the crowds dynamics for time-dependent reputation and ranking. In *In ASONAM '09: Proceedings of the International Conference on Advances in Social Network Analysis and Mining*, July 2009.

[GGLNT04] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA, 2004. ACM.

[HRW08] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *CoRR*, abs/0812.1045, 2008.

[JSFT07] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[JZSC09] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Microblogging as online word of mouth branding. In *CHI '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 3859–3864, New York, NY, USA, 2009. ACM.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[Moo10] Robert J. Moore. New data on twitters users and engagement. `http://tinyurl.com/ybh2fa9`, 2010.

[Sem10] Semiocast. Half of messages on twitter are not in english. `http://tinyurl.com/3a9pcp4`, 2010.

[SYC09] Nishanth Sastry, Eiko Yoneki, and Jon Crowcroft. Buzztraq: predicting geographical access patterns of social cascades using social networks. In *SNS '09: Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, pages 39–45, New York, NY, USA, 2009. ACM.

[ZR09] Dejin Zhao and Mary Beth Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *GROUP '09: Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252, New York, NY, USA, 2009. ACM.

Appendix A

Terms and Users Used to Acquire Election Data

Note that the Twitter API does not accept '#' symbols in filter terms, so there is no distinction between hashtags and topics; I have used the distinction in this table to show which words were being used mainly as hashtags during the election.

Topics	Hashtags	Users
#anyonebutcameron	BNP	@BBCElection
#cleggy	Labour	@Conservatives
#davidcameron	Clegg	@EricPickles
#dcameron	Libdem	@LabourParty
#electionday	Tories	@LibDems
#GE10	Tory	@Nick_Clegg
#gordonbrown	UKIP	@UKLabour
#greenparty		
#hangem		
#imnotvotingconservative		
#imnotvotinglabour		
#imvotingconservative		
#imvotinglabour		
#ldem		
#libdems		
#nickclegg		
#philippastroud		
#toryvote		
#torywin		
#ukvote		
#ukelection		

Table A.1: Terms used to filter election data

Appendix B

Users with known party affiliation

Username	Party	Real Name	Type
Conservatives	Conservative	N/A	Party
EricPickles	Conservative	Eric Pickles	Politician
henrymacrory	Conservative	Henry Macrory	Party Employee
HMSEnterprise	Conservative	Shane McMurray	Blogger
SamuelCoates	Conservative	Samuel Coates	Party Empoyee
campbellclaret	Labour	Alastair Campbell	Politician
eddieizzard	Labour	Eddie Izzard	Celebrity
johnprescott	Labour	John Prescott	Politician
LabourList	Labour	N/A	Party
LabourParty	Labour	N/A	Party
tom_watson	Labour	Tom Watson	Politician
UKLabour	Labour	N/A	Party
UKLabourParty	Labour	N/A	Party
libdems	Lib Dem	N/A	Party
Nick_Clegg	Lib Dem	Nick Clegg	Politician
stevebeasant	Lib Dem	Steve Beasant	Politician

Table B.1: Seeds for party affiliation propagation

Appendix C

Election Graphs

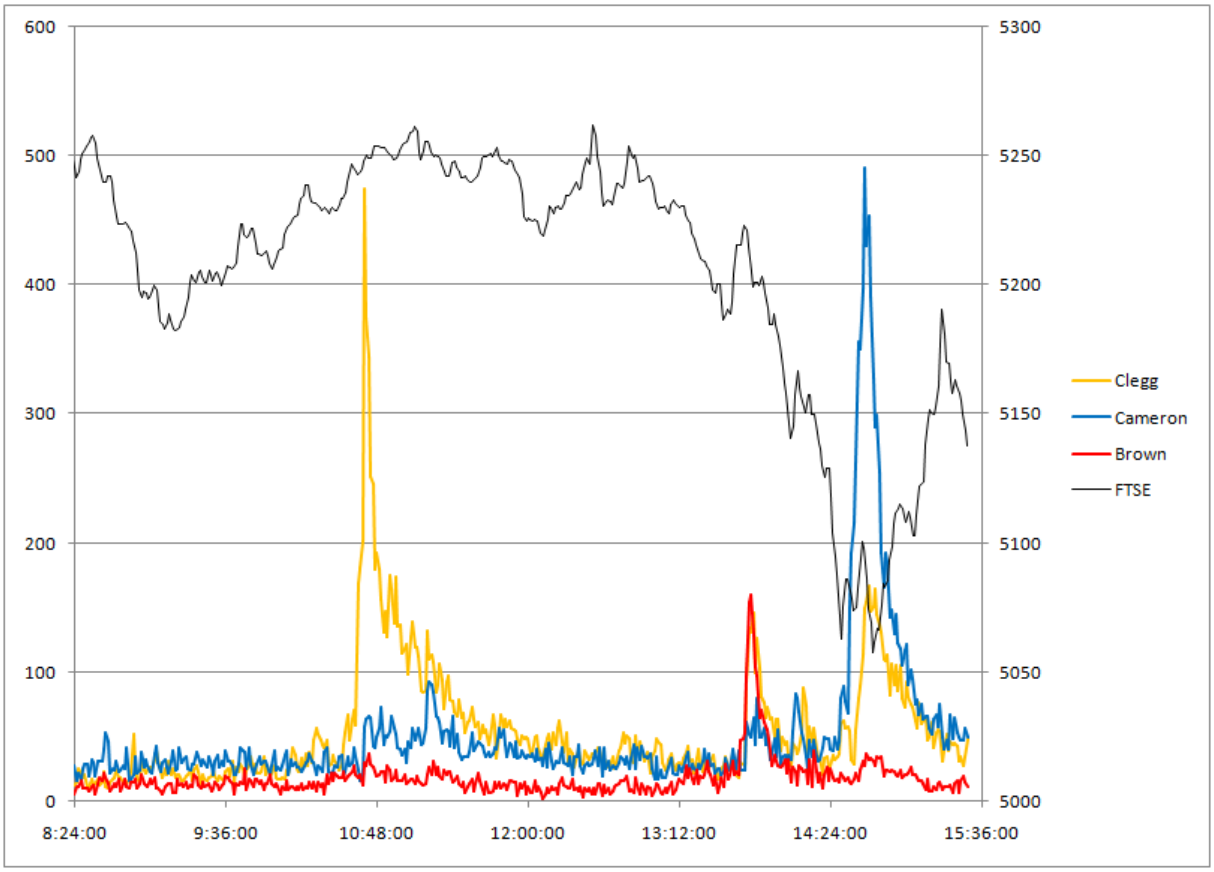


Figure C.1: Party leader mentions and the FTSE 100 Index during the day after around the election

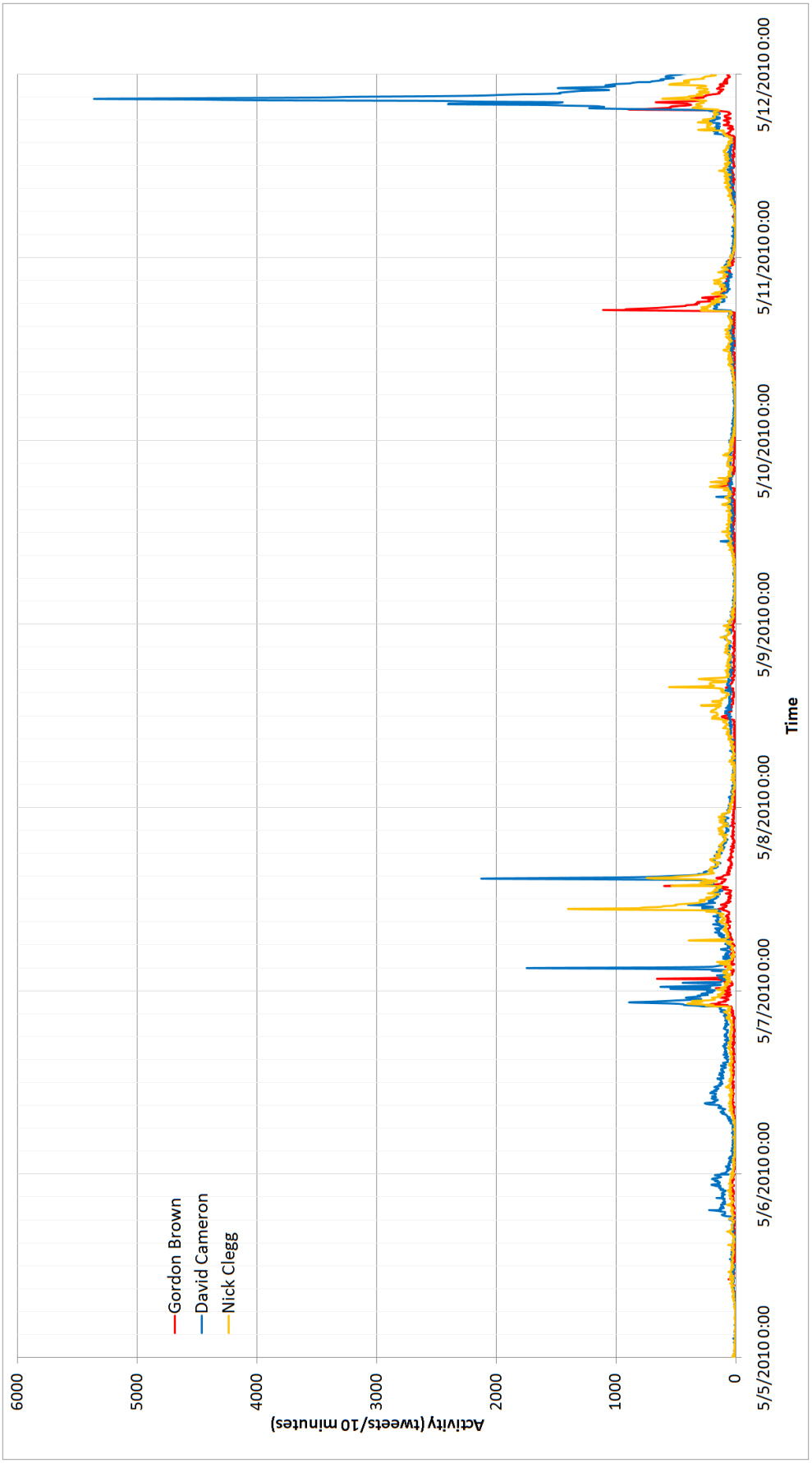


Figure C.2: Party leader mentions during the week around the election

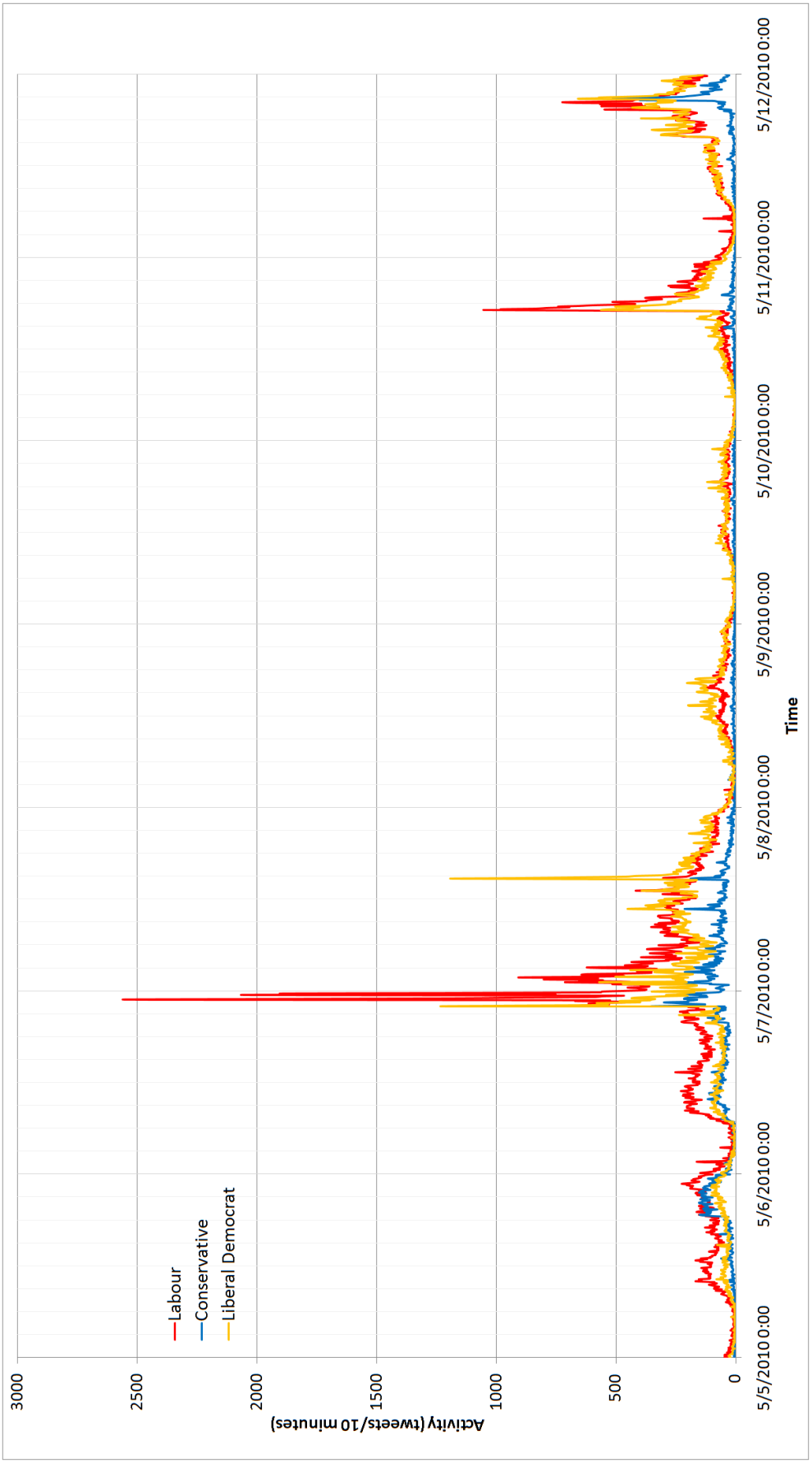


Figure C.3: Party mentions during the week around the election

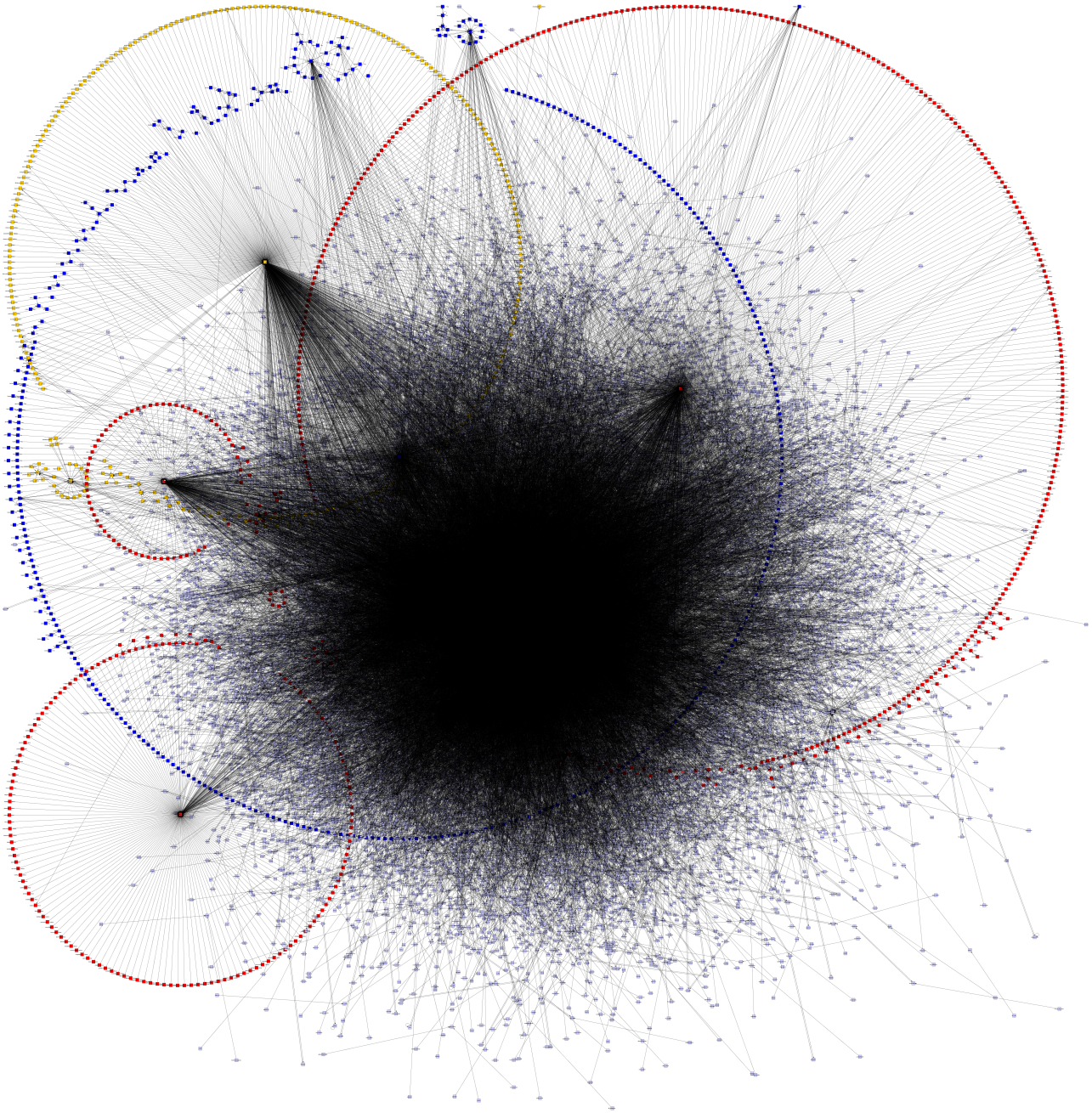


Figure C.4: Graph colored by party affiliation from a 157,764 tweet sample from 05:43 7/5/2010 to 23:17 11/5/2010

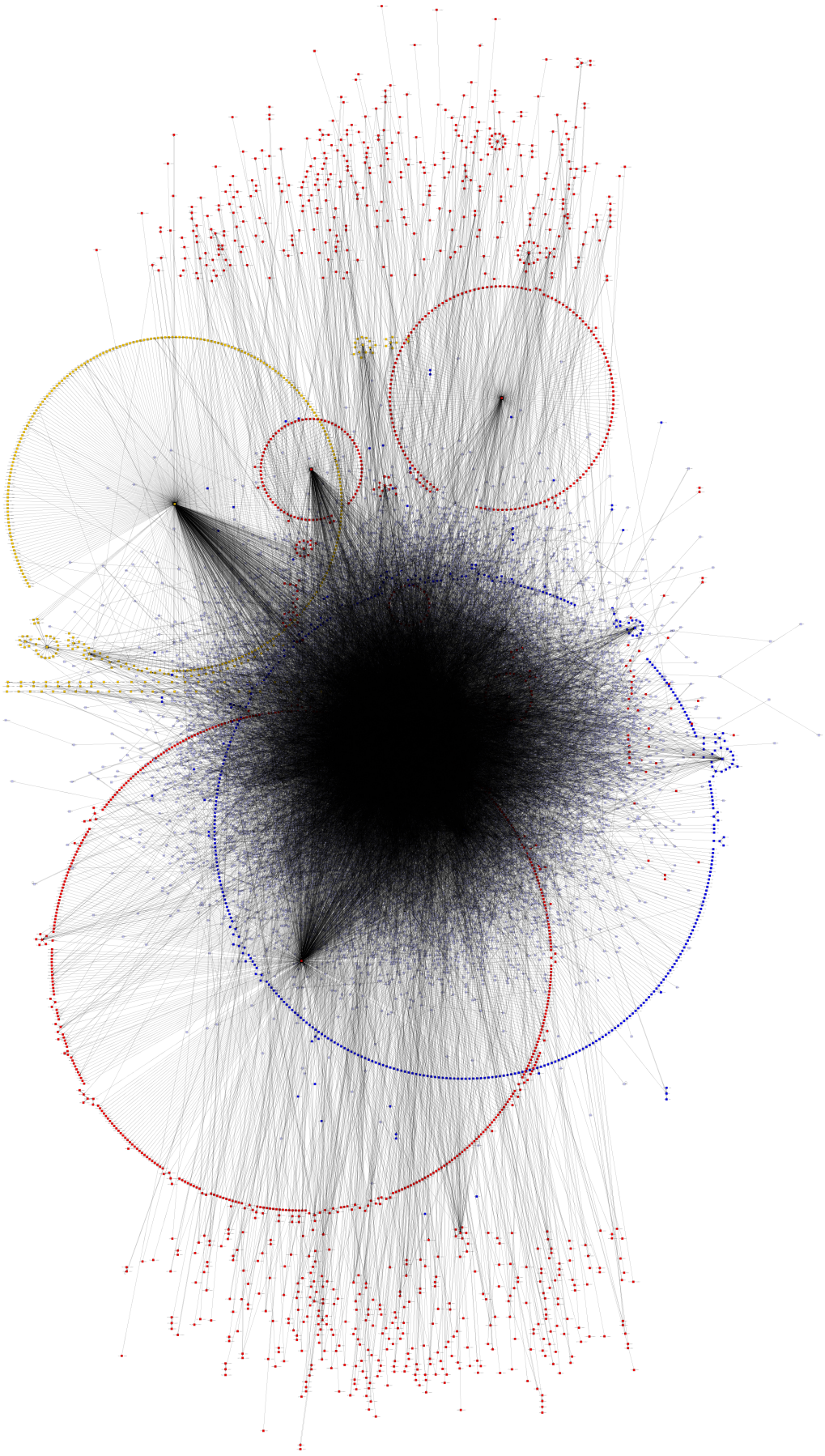


Figure C.5: Graph colored by party affiliation from a 157,764 tweet sample from 05:43 7/5/2010 to 23:17 11/5/2010

Appendix D

Trending Topics

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
#2pmforever7	13046	0.065	0.323	1.000	0.049
#6wordstory	10765	0.061	0.202	1.000	0.012
#agqr	17187	0.026	0.078	1.000	0.013
#ahaLN	10735	0.109	0.156	1.000	0.078
#Ahater	11687	0.066	0.274	1.000	0.032
#aibou	25521	0.045	0.290	1.000	0.031
#alottayall	10043	0.037	0.249	1.000	0.168
#americanidol	24635	0.083	0.098	1.000	0.170
#amitheonlyone	71725	0.060	0.256	1.000	0.077
#AndThenWeHadSex	19025	0.075	0.269	1.000	0.039
#areyoukiddingme	18936	0.062	0.128	1.000	0.329
#AreYouStupid	21098	0.062	0.261	1.000	0.030
#AstonsTattoo	18237	0.267	0.291	1.000	0.026
#AwardGoes2	13031	0.221	0.416	1.000	0.021
#awesomeindianthings	21686	0.080	0.197	1.000	0.125
#AwHellNah	17806	0.075	0.305	1.000	0.053
#badgirlsclub	22300	0.108	0.169	1.000	0.086
#bb0407	21040	0.023	0.062	1.000	0.834
#bbcqt	34156	0.111	0.281	1.000	0.034
#beastshock	24526	0.116	0.149	1.000	0.045
#BeforeIdie I	19012	0.140	0.309	1.000	0.012
#BETmessedUpWhen	98716	0.044	0.345	1.000	0.039

Table D.1: Trending topics with over 10,000 tweets (1)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
#Call1800FML	22833	0.062	0.264	1.000	0.010
#CBOnRadio	23315	0.148	0.247	1.000	0.022
#CelebrityApprentice	30908	0.217	0.183	1.000	0.026
#ChancesAre	29151	0.078	0.221	1.000	0.113
#ChileNeedsJonas	20861	0.178	0.309	1.000	0.112
#ChileWantsJonas	23050	0.130	0.461	1.000	0.009
#chipmunkfan	17959	0.412	0.285	1.000	0.044
#chucknorris	24269	0.099	0.257	1.000	0.062
#ClockOut	28221	0.074	0.360	1.000	0.046
#Dead	17426	0.099	0.379	1.000	0.043
#debill	37161	0.159	0.415	1.000	0.422
#dontcallyourself	36520	0.040	0.304	1.000	0.073
#dontgotogether	14948	0.046	0.277	1.000	0.065
#Eastenders	65054	0.082	0.163	1.000	0.028
#elimakesmetingle	10043	0.222	0.129	1.000	0.076
#EverFeelLike	41199	0.040	0.282	1.000	0.042
#everydayiwakeup	45580	0.056	0.255	1.000	0.062
#fabsmixtape	20933	0.115	0.608	1.000	0.074
#faktanya	47646	0.035	0.609	1.000	0.046
#fallinginlove	29999	0.078	0.283	1.000	0.037
#fatorwhore	10038	0.343	0.308	1.000	0.018
#FF	353184	0.639	0.294	1.000	0.067
#firstdaterules	27458	0.033	0.272	1.000	0.063
#FollowFriday	227963	0.639	0.293	1.000	0.064
#followmeJP	19148	0.111	0.290	1.000	0.154
#followmeliltwist	12739	0.624	0.031	1.000	0.177
#FollowSaturday	13898	0.191	0.728	1.000	0.054
#followsunday	18985	0.203	0.511	1.000	0.262
#followtuesday	10472	0.243	0.676	1.000	0.063
#forUs	15464	0.029	0.889	1.000	0.080
#FuckYourOpinion	11396	0.055	0.272	1.000	0.132
#grandesduos	15215	0.129	0.280	1.000	0.062
#grandmawhy	18114	0.047	0.183	1.000	0.123
#greasyleader	24756	0.127	0.271	1.000	0.079

Table D.2: Trending topics with over 10,000 tweets (2)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
#HappyBdayGerardWay	19995	0.096	0.124	1.000	0.377
#HappyBdayRihanna	15497	0.071	0.388	1.000	0.075
#HappyBirthdayJustin	46790	0.254	0.286	1.000	0.039
#hardwithoutshoes	13164	0.166	0.319	1.000	0.152
#hbu	72906	0.028	0.702	1.000	0.060
#hcr	43756	0.130	0.511	1.000	0.461
#HCRsummit	12255	0.075	0.409	1.000	0.174
#heroes100	19632	0.067	0.645	1.000	0.043
#hitsunami	15055	0.070	0.377	1.000	0.386
#honorsocietytour	12617	0.419	0.191	1.000	0.183
#howuathug	12147	0.049	0.224	1.000	0.021
#howwouldyoufeel	17304	0.069	0.238	1.000	0.059
#HowYouAManBut	94412	0.035	0.258	1.000	0.019
#howyouathug	31100	0.043	0.243	1.000	0.026
#HumanoidCityTour	38481	0.113	0.237	1.000	0.055
#HumanoidCityTourTH	14537	0.118	0.187	1.000	0.036
#Ibelieve	60759	0.088	0.242	1.000	0.148
#icheatedbecause	57466	0.041	0.228	1.000	0.078
#iDoit2	100249	0.014	0.868	1.000	0.032
#ifyourdominican	11827	0.053	0.359	1.000	0.045
#ifyourpuertorican	16838	0.051	0.357	1.000	0.039
#igotplayed	23719	0.253	0.145	1.000	0.015
#ihatequotes	87094	0.027	0.909	1.000	0.032
#iHeardChuckNorris	78324	0.077	0.298	1.000	0.022
#iicdhouse	10073	0.150	0.107	1.000	0.013
#ijustwannathank	30171	0.202	0.226	1.000	0.038
#iLoveFAMU	84944	0.125	0.455	1.000	0.099
#ILoveItWhenTrey	30898	0.146	0.232	1.000	0.032
#ILoveLegacyCuz	17038	0.505	0.129	1.000	0.086
#ILoveLilTwistCuz	18673	0.389	0.190	1.000	0.006

Table D.3: Trending topics with over 10,000 tweets (3)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
#imathug	26925	0.084	0.213	1.000	0.057
#imattractedto	35687	0.071	0.227	1.000	0.121
#imcurious	11955	0.059	0.212	1.000	0.187
#imfromteam	14874	0.068	0.348	1.000	0.028
#imnotattractedto	20048	0.029	0.161	1.000	0.233
#ImNotBeingRudeBut	36426	0.054	0.223	1.000	0.056
#imtiredof	38737	0.040	0.239	1.000	0.099
#Imtiredofseeing	75588	0.040	0.286	1.000	0.043
#INeverWannaFeelThat	14335	0.032	0.213	1.000	0.121
#infolimit	38112	0.175	0.806	1.000	0.041
#inhighschool I	19009	0.103	0.279	1.000	0.045
#inmyfamily	11789	0.048	0.190	1.000	0.017
#iREFUSEto	64250	0.057	0.245	1.000	0.042
#iStock10	12720	0.284	0.039	1.000	0.381
#its2010whyyoustill	52787	0.040	0.236	1.000	0.086
#ItsFunnyHow	57721	0.045	0.219	1.000	0.188
#itsnotcool	29550	0.066	0.270	1.000	0.063
#ItsNotOkay	32642	0.058	0.287	1.000	0.038
#iWillNever	40703	0.090	0.291	1.000	0.069
#iwishyouwouldstop	29880	0.060	0.221	1.000	0.038
#iwonderhow	12272	0.066	0.220	1.000	0.045
#javamusikindo	13850	0.794	0.133	1.000	0.057
#JavaRockingland2010	71408	0.192	0.515	1.000	0.048
#jedwardpic	14294	0.691	0.086	1.000	0.219
#jesseCHAT	46051	0.258	0.089	1.000	0.023
#JonasAreBack	32811	0.136	0.290	1.000	0.035
#JonasBrothersAre	22697	0.056	0.355	1.000	0.030
#JonasBrothersAre.bk	22697	0.056	0.355	1.000	0.030
#JonasWorldTour2010	16003	0.127	0.451	1.000	0.031
#jorts	19859	0.291	0.233	1.000	0.062
#justcausewecool	21965	0.052	0.195	1.000	0.052
#justice4MJ	23975	0.098	0.423	1.000	0.080
#justinbiebermyspace	13082	0.039	0.022	1.000	0.963

Table D.4: Trending topics with over 10,000 tweets (4)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
#KandiOnUstream	36583	0.044	0.072	1.000	0.970
#KBShow	55997	0.682	0.127	1.000	0.030
#KeriHilsonOnUstream	15769	0.042	0.056	1.000	0.880
#KingstonFollows	85227	0.868	0.100	1.000	0.005
#kissmesoohyun	13342	0.230	0.139	1.000	0.088
#KushWillMake	12224	0.067	0.271	1.000	0.044
#LetsBeReal	42206	0.082	0.264	1.000	0.078
#LilTwistTakeover	15843	0.227	0.164	1.000	0.551
#LilWaynesBreath	11732	0.055	0.229	1.000	0.249
#LilWaynesNextExcuse	92410	0.026	0.316	1.000	0.337
#MadeInChina	24431	0.068	0.285	1.000	0.026
#MarchMadness	12063	0.109	0.125	1.000	0.107
#marchwish	30087	0.067	0.422	1.000	0.055
#MechanicalDummy	10645	0.109	0.234	1.000	0.016
#MeWithoutYouIsLike	93770	0.093	0.197	1.000	0.046
#mm	89182	0.200	0.183	1.000	0.295
#mubankSNSD	11240	0.051	0.237	1.000	0.034
#musicmonday	109027	0.212	0.153	1.000	0.389
#Mylifeasliz	27610	0.057	0.104	1.000	0.556
#NewRule	12760	0.051	0.339	1.000	0.077
#nooffense	37778	0.096	0.230	1.000	0.031
#NothingWorseThan	23717	0.047	0.234	1.000	0.072
#NotMeThough	14338	0.043	0.256	1.000	0.022
#nowplaying	665717	0.117	0.095	1.000	0.194
#OhJustLikeMe	129909	0.023	0.856	1.000	0.084
#OkJokesOver	31842	0.074	0.230	1.000	0.040
#omgfacts	26838	0.044	0.734	1.000	0.196
#omgthatssotrue	27414	0.038	0.834	1.000	0.017
#OMJretweetif	28352	0.022	0.815	1.000	0.085
#OnaScaleFrom 1	56860	0.119	0.303	1.000	0.011
#OnAScaleFrom	17682	0.116	0.296	1.000	0.015
#Onlyyoushawty	18474	0.120	0.615	1.000	0.009

Table D.5: Trending topics with over 10,000 tweets (5)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
#PictureThat	24196	0.094	0.190	1.000	0.086
#PlacesIWannaHaveSex	84614	0.066	0.313	1.000	0.047
#PMQs	11392	0.087	0.296	1.000	0.085
#PrayForTheWorld	10340	0.040	0.348	1.000	0.028
#PROMOTE RT BUAT	13617	0.004	0.911	1.000	0.054
#raiseyourhandif	53196	0.065	0.369	1.000	0.053
#Randomthought	33436	0.070	0.182	1.000	0.038
#RelationshipRules	99704	0.029	0.384	1.000	0.043
#relationshiptips	18857	0.029	0.389	1.000	0.044
#ReTweetThis	17251	0.054	0.618	1.000	0.053
#Rhamu	10959	0.249	0.288	1.000	0.033
#RIPAlejandraJonas	36259	0.085	0.532	1.000	0.051
#ripbig	51385	0.072	0.280	1.000	0.096
#RIPLaylaGrace	11377	0.167	0.297	1.000	0.063
#saveBBC6music	15016	0.100	0.417	1.000	0.296
#SelenaGomezLive	16063	0.032	0.027	1.000	0.842
#SexualAttractions	43846	0.066	0.331	1.000	0.033
#shootup	55155	0.021	0.084	1.000	0.051
#SimonSays	14515	0.125	0.217	1.000	0.048
#sincewhen	19093	0.052	0.272	1.000	0.029
#SkeeterPanLuhLike	24186	0.247	0.211	1.000	0.023
#SomewhereRightNow	126067	0.062	0.244	1.000	0.038
#SoProudOfYouNickJ	76794	0.218	0.288	1.000	0.201
#souljaboytellem	21956	0.343	0.128	1.000	0.344
#ss2shanghai	10034	0.109	0.408	1.000	0.080
#Stay	30990	0.239	0.235	1.000	0.205
#strippernames	21028	0.090	0.287	1.000	0.013
#SulHee	27308	0.129	0.268	1.000	0.039
#SummerRules	33271	0.039	0.283	1.000	0.039
#TDL	19867	0.016	0.927	1.000	0.026
#TEDxNYED	18228	0.265	0.451	1.000	0.162
#TelephoneVideo	11226	0.116	0.293	1.000	0.077
#TextDatGetULockedUp	10946	0.043	0.289	1.000	0.018
#textsihate	35630	0.028	0.261	1.000	0.050

Table D.6: Trending topics with over 10,000 tweets (6)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
#thatssoannoying	77641	0.037	0.197	1.000	0.237
#ThatsWhyILeftYou	16661	0.030	0.249	1.000	0.033
#THB	27602	0.014	0.914	1.000	0.065
#theburiedlife	18836	0.270	0.150	1.000	0.060
#therealme	24358	0.044	0.221	1.000	0.080
#THfansRock	13580	0.113	0.250	1.000	0.019
#thingsCHEATERSsay	41118	0.031	0.250	1.000	0.022
#ThingsUglyPeopleSay	18711	0.036	0.256	1.000	0.020
#ThingsWeWantBack	66164	0.050	0.310	1.000	0.038
#thisismydream	10128	0.145	0.304	1.000	0.103
#ThrowBackLyrics	46904	0.032	0.098	1.000	0.228
#TLS	113308	0.014	0.913	1.000	0.061
#TomKaulitzSexTape	26852	0.109	0.268	1.000	0.036
#tosavemoney	11590	0.045	0.193	1.000	0.135
#toysoldiers	35453	0.714	0.075	1.000	0.138
#ttblogtv	20605	0.104	0.485	1.000	0.030
#tweetorangkaya	15889	0.062	0.334	1.000	0.052
#tweetsIDGAFabout	59828	0.047	0.325	1.000	0.046
#twitterislike	72903	0.036	0.299	1.000	0.167
#ugotmefkdup	10021	0.052	0.213	1.000	0.020
#UJustMadeItWorse	55302	0.049	0.192	1.000	0.023
#UKnowUBroke	106679	0.053	0.227	1.000	0.070
#UKnowUrHigh	13239	0.051	0.247	1.000	0.133
#UNotFromTheHoodif	107105	0.037	0.256	1.000	0.089
#UrParentsEver	23797	0.026	0.277	1.000	0.040
#urwack	21563	0.081	0.250	1.000	0.080
#Ustream@SXSW	18885	0.996	0.004	1.000	0.991
#van2010	16512	0.059	0.266	1.000	0.118
#WeAdoreDemi	23470	0.190	0.343	1.000	0.034
#weAdoreJustin	27118	0.095	0.490	1.000	0.018
#WeGoTogetherLike	70890	0.072	0.195	1.000	0.105
#WeLoveMiley	16786	0.062	0.435	1.000	0.009
#weloveselena	18747	0.142	0.286	1.000	0.018

Table D.7: Trending topics with over 10,000 tweets (7)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
#WeSupportTokioHotel	22023	0.123	0.256	1.000	0.025
#whatifGod	10296	0.047	0.246	1.000	0.059
#Whatsthebigdeal	41196	0.089	0.233	1.000	0.041
#WhatWouldYouRather	54991	0.048	0.327	1.000	0.045
#WhoIsLilTwist	21410	0.272	0.294	1.000	0.130
#WhoLiedToYou	13489	0.053	0.200	1.000	0.026
#WhyPeopleOnTwitter	10320	0.039	0.360	1.000	0.077
#whyursingle	24765	0.047	0.278	1.000	0.030
#whyyomama	14485	0.065	0.155	1.000	0.040
#whyyoursingle	10424	0.037	0.244	1.000	0.019
#WorldEvanescenceDay	21907	0.149	0.095	1.000	0.034
#Yeaisaidit	16887	0.076	0.280	1.000	0.030
#Yotwit	70019	0.007	0.015	1.000	0.912
#youaintforme	20836	0.037	0.216	1.000	0.034
#youdidntwantmeuntil	10410	0.027	0.175	1.000	0.018
#youknowitslovewhen	78017	0.047	0.229	1.000	0.208
#youmightwannastop	12499	0.052	0.239	1.000	0.035
#YoureFIRED	27551	0.061	0.356	1.000	0.041
#YourFaceMakesMe	17352	0.091	0.214	1.000	0.030
#ZodiacFacts	31845	0.024	0.872	1.000	0.076
A New Meme	13377	0.089	0.231	0.115	0.861
ABDC	17384	0.117	0.103	0.372	0.036
Actor Corey Haim	14463	0.035	0.487	0.078	0.649
Adam Lambert	30553	0.174	0.331	0.100	0.218
Alex Lambert	10777	0.173	0.229	0.142	0.093
Alice In Wonderland	105361	0.137	0.126	0.047	0.127
Alice	287204	0.189	0.149	0.089	0.191
Amen'	11632	0.381	0.430	0.159	0.103
American Idol	77980	0.133	0.124	0.101	0.240
Andrew Koenig	14026	0.041	0.476	0.166	0.640
ANTM	11690	0.152	0.115	0.477	0.024
Apple iPad	30187	0.199	0.410	0.235	0.888
Arsenal	27835	0.149	0.212	0.185	0.257

Table D.8: Trending topics with over 10,000 tweets (8)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
Avatar	30722	0.289	0.171	0.192	0.259
Bachelor	10623	0.164	0.154	0.338	0.042
Bad Girls Club	12217	0.166	0.164	0.126	0.043
BAFTAs	16932	0.095	0.182	0.609	0.176
BBC	49519	0.102	0.235	0.315	0.646
Bears	10960	0.177	0.189	0.165	0.356
Betty White	14508	0.080	0.378	0.241	0.471
Beyonce	30274	0.187	0.203	0.194	0.363
Beyonce?	15047	0.171	0.193	0.181	0.402
BGC	14213	0.156	0.201	0.630	0.027
Big Mike	10229	0.217	0.108	0.159	0.121
Biggie	46541	0.163	0.277	0.388	0.146
Calm	13611	0.314	0.106	0.073	0.275
Canada	52910	0.168	0.158	0.210	0.236
Canadian	11291	0.154	0.210	0.177	0.255
CERN	10735	0.207	0.305	0.267	0.439
Champions League	10943	0.070	0.165	0.118	0.502
Chelsea	34190	0.135	0.218	0.191	0.178
Chile	222667	0.175	0.335	0.293	0.377
Chris Brown	16887	0.172	0.313	0.316	0.205
Chuck Norris	72156	0.171	0.317	0.197	0.094
CNN	10132	0.169	0.390	0.202	0.346
CODY	22607	0.383	0.299	0.085	0.158
Corey Haim	35967	0.069	0.173	0.073	0.394
Could Netflix	13632	0.113	0.289	0.083	0.791
Currently Fair	10477	0.006	0.005	0.006	0.942
CURRENTLY Partly	23483	0.003	0.010	0.023	0.627
Dear Terrorist	38432	0.052	0.445	0.134	0.042
Drake	26112	0.198	0.216	0.238	0.217
EastEnders	16312	0.142	0.135	0.346	0.031
Easter	41465	0.208	0.124	0.110	0.454
Eclipse	67038	0.133	0.211	0.186	0.375
Eenie Meenie	42639	0.379	0.305	0.147	0.105

Table D.9: Trending topics with over 10,000 tweets (9)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
Everton	10555	0.119	0.189	0.183	0.139
Famu	17671	0.252	0.281	0.418	0.088
Felicia Anjani	18103	0.026	0.831	0.788	0.033
Finland	12908	0.091	0.169	0.258	0.113
Follow Friday	53820	0.585	0.241	0.386	0.101
Foursquare	11126	0.466	0.234	0.103	0.550
Georgetown	13963	0.127	0.156	0.207	0.137
Gimana	11234	0.296	0.426	0.050	0.114
Glee	44767	0.214	0.142	0.284	0.106
Go Canada Go	18731	0.117	0.148	0.162	0.040
Gold	17649	0.117	0.244	0.181	0.207
Goodmorning	74485	0.234	0.126	0.106	0.058
Goodnight	258454	0.230	0.053	0.105	0.055
Google Maps	14919	0.061	0.383	0.246	0.822
Gossip Girl	11134	0.122	0.147	0.123	0.137
GQ95z6ywcBY	12614	0.202	0.119	0.112	0.947
Haiti	142521	0.206	0.299	0.246	0.472
Happy Womens Day	26236	0.152	0.276	0.104	0.146
Happy Women's Day	39073	0.148	0.278	0.106	0.141
Hawaii	78015	0.099	0.440	0.250	0.294
High	35098	0.204	0.121	0.119	0.397
HTC	25305	0.072	0.172	0.219	0.817
Hurt Locker	31545	0.158	0.191	0.120	0.260
i think im pregnant	22976	0.060	0.398	0.271	0.115
Im Back	17161	0.295	0.136	0.126	0.131
IHOP	42635	0.220	0.206	0.180	0.137
I'm Back	16920	0.297	0.125	0.119	0.116
Indonesia	38960	0.144	0.440	0.128	0.198
INDONESIAN ELFs	12071	0.112	0.449	0.511	0.200
Indonesian Idol	16988	0.088	0.535	0.095	0.100
iPad	28350	0.141	0.149	0.458	0.827
iPhone OS 4	38909	0.054	0.234	0.128	0.679
IPL	20334	0.171	0.102	0.401	0.268

Table D.10: Trending topics with over 10,000 tweets (10)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
Ireland	11281	0.196	0.098	0.146	0.386
Jaebum	14474	0.068	0.322	0.455	0.109
Jake	14224	0.141	0.124	0.209	0.069
Japan	30601	0.117	0.395	0.200	0.427
Java Rockin'Land	62410	0.149	0.798	0.090	0.104
Jemi	25591	0.257	0.287	0.156	0.124
JJF	13411	0.266	0.441	0.183	0.153
Jonas	29903	0.231	0.358	0.318	0.126
JonasInArgentina2010	27878	0.138	0.289	0.190	0.107
Jonghyun	19806	0.107	0.184	0.051	0.043
JR Smith	10139	0.142	0.180	0.146	0.081
JUST RT ASAP	16130	0.042	0.918	0.511	0.089
Justin Bieber	937528	0.150	0.231	0.183	0.358
Kathryn Bigelow	10774	0.057	0.243	0.163	0.359
KKR	11040	0.295	0.091	0.277	0.091
KNBC	25256	0.056	0.876	0.031	0.061
Kobe	14371	0.271	0.248	0.214	0.109
Ladies	17903	0.214	0.407	0.197	0.214
Lady Gaga	190610	0.146	0.184	0.189	0.376
Lakers	26678	0.211	0.223	0.226	0.113
Law	15489	0.161	0.256	0.134	0.423
Lil Wayne	26379	0.153	0.310	0.221	0.346
Lost Boys	19502	0.200	0.167	0.139	0.252
Low	10148	0.210	0.133	0.145	0.381
Mac Heist	11804	0.007	0.009	0.004	0.983
Mac	20208	0.228	0.140	0.107	0.452
Malcolm McLaren	11968	0.038	0.311	0.092	0.591
March Madness	30687	0.119	0.115	0.117	0.352
Martin Skoula	12719	0.058	0.450	0.203	0.230
MN Winds	10207	0.001	0.002	0.000	0.955
Moscow	11680	0.058	0.262	0.230	0.639
My World 2	90103	0.459	0.349	0.089	0.053
Name Ya Top 5 Biggie	13259	0.063	0.587	0.177	0.089

Table D.11: Trending topics with over 10,000 tweets (11)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
Naomi Campbell	10611	0.052	0.293	0.092	0.516
National Grammar Day	11636	0.097	0.417	0.255	0.372
NBC	10330	0.219	0.298	0.340	0.149
NCAA	60187	0.101	0.132	0.291	0.467
Never Let You Go	34189	0.406	0.364	0.343	0.125
New Moon	21452	0.142	0.105	0.078	0.146
Ohio	10800	0.173	0.178	0.228	0.220
Olympics	29862	0.146	0.149	0.379	0.238
OMGDUVALFACT	12120	0.036	0.904	0.059	0.054
Oscars	154937	0.147	0.198	0.238	0.432
Pacific Tsunami	17543	0.043	0.591	0.387	0.467
Pacific	13892	0.074	0.521	0.327	0.452
Paramore	15086	0.165	0.300	0.360	0.055
Perhatian	13347	0.184	0.485	0.109	0.145
PlayStation Move	18525	0.041	0.136	0.151	0.764
Pocong	11055	0.264	0.360	0.593	0.068
PROMOTE ONE BY ONE	20749	0.010	0.920	0.136	0.044
PROMOTE PART 2	14987	0.057	0.806	0.028	0.142
PROMOTE SORE ADA	12638	0.024	0.738	0.044	0.048
Promote	27526	0.127	0.586	0.153	0.225
PROMOTESORE	13112	0.025	0.849	0.181	0.036
QVC	22063	0.423	0.292	0.043	0.164
RETWEET THIS IF YOU	31524	0.038	0.677	0.257	0.137
Ricky Martin	19903	0.095	0.299	0.099	0.219
RIP Corey Haim	10062	0.028	0.164	0.095	0.168
RT 50 Orang	14152	0.018	0.803	0.259	0.107
RT CUMA 10 MENIT	33006	0.027	0.631	0.165	0.045
RT IF YOU	19307	0.038	0.848	0.296	0.105
RT RT RT	34547	0.033	0.948	0.224	0.249
Sachin	11114	0.098	0.163	0.310	0.074
Sandra Bullock	19238	0.114	0.196	0.130	0.306
SBY	11893	0.110	0.459	0.080	0.227
SCTV NOW	21404	0.131	0.549	0.067	0.079

Table D.12: Trending topics with over 10,000 tweets (12)

Topic	Tweets	Mention Rate	Retweet Rate	Hashtag Rate	URL Rate
Seuss	22548	0.074	0.338	0.197	0.349
Shutter Island	81199	0.138	0.070	0.236	0.279
siapa aja boleh	17983	0.067	0.519	0.102	0.085
Skins	38696	0.188	0.081	0.231	0.068
Smile	17646	0.246	0.298	0.124	0.143
SNL	13215	0.144	0.139	0.378	0.077
Spring	14756	0.169	0.121	0.109	0.336
SXSW	25222	0.327	0.233	0.532	0.460
Team USA	14874	0.095	0.224	0.171	0.107
Telephone	63355	0.193	0.184	0.197	0.308
Texas	12097	0.176	0.159	0.175	0.388
TGIF	70002	0.151	0.120	0.244	0.105
There Is No	17626	0.274	0.192	0.142	0.220
This You	16363	0.313	0.256	0.110	0.419
THU Rain	12725	0.019	0.046	0.040	0.304
TNA	19653	0.271	0.119	0.356	0.093
Toyota	11543	0.076	0.204	0.200	0.594
Tron Legacy	12706	0.057	0.266	0.125	0.772
Tsunami	95550	0.103	0.428	0.237	0.403
TUE Rain	14039	0.034	0.041	0.040	0.265
Turkey	13578	0.131	0.247	0.123	0.458
TvRock	21249	0.147	0.001	0.145	0.002
U Smile	19439	0.493	0.249	0.264	0.119
USA	43301	0.168	0.149	0.205	0.207
Vancouver	14877	0.078	0.352	0.257	0.552
ViagPure	13992	0.019	0.007	0.009	0.954
Vienna	18810	0.171	0.127	0.230	0.045
Washington	16997	0.074	0.141	0.197	0.658
WeNeedJonasInArg	30296	0.242	0.278	0.197	0.082
WIND	37081	0.113	0.047	0.172	0.204
Without God	13591	0.045	0.401	0.190	0.061
Wonderland	12943	0.163	0.247	0.067	0.235
WWE	11480	0.269	0.118	0.397	0.150
YANG ANAK INDONESIA	90757	0.032	0.618	0.158	0.066
YANG MAU GUA	19027	0.077	0.661	0.240	0.048
YES WE WANT PITBULL	18565	0.031	0.815	0.749	0.129

Table D.13: Trending topics with over 10,000 tweets (13)