Inferring Interestingness in Online Social Networks

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

William M. Webberley

May 2014

Cardiff University
School of Computer Science & Informatics

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date ....

Statement 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed .... (candidate)

Date ....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed .... (candidate)

Date ....

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date ....

To my family and friends;
To George, Pete,
Jack, Max, Ella,
Tom, and Charlie.

This is for their support.

Acknowledgements

This thesis and the research it contains could not exist if not for the unfailing support of my supervisors, Roger Whitaker and, particularly, Stuart Allen. Their guidance and input have provided drive, shaped my research and have given me the confidence required for producing and defending ideas both within and outside of research. My utmost gratitude is extended to them for this and for their continual encouragement at all stages of my PhD, always with fresh perspective and clarification for research pathways and for keeping me on track. I feel confident in saying that I was very lucky in being in their supervision and I would not be the same person, professionally and otherwise, today if it wasn't for their combined inputs over the past few years.

I'd also like to thank Martin Chorley for his continuously helpful and enthusiastic crowdsourcing expertise, without which the research in this thesis could not have been validated in the way it has been. Matt Williams has provided insight throughout my time as a PhD student, from long sessions discussing the tiniest irrelevant details to recommendations on research direction and scope. Discussions with him have helped form many of my research ideas. I am also deeply grateful to Chris Gwilliams for his helpful input over course of my PhD, including information on where one might find the most cost-effective M.O.T. deals through to how best to structure complex and fiddly data queries. Thanks also to Matt John, who, ever critical, helped proof-read this thesis and provided invaluable constructive feedback on notation and nomenclature.

My family and close friends have always pushed me and provided encouragement to make me who I am today. If not for them, I would not have had the confidence to carry out and defend my research. They deserve my utmost thanks.

Finally, my time as a research student would not have been the same without the constant presence and support from the School of Computer Science & Informatics, the other members of the MobiSoc group, the 'superteam', and non-academic friends. In addition to those named previously, Gualtiero Colombo, Ian Cooper, Jon Quinn, Diego Pizzocaro, Rich Coombs, Liam Turner, Nick Sharp, and Ross Taylor have made the time spent researching my PhD both interesting and enjoyable and I am extremely grateful to them.

Abstract

Information sharing and user-generated content on the Internet has given rise to the increased presence of uninteresting and 'noisy' information in media streams on many online social networks. Although there is a lot of 'interesting' information also shared amongst users, the noise increases the cognitive burden in terms of the users' abilities to identify what is interesting and may increase the chance of missing content that is useful or important. Additionally, users on such platforms are generally limited to receiving information only from those that they are directly linked to on the social graph, meaning that users exist within distinct content 'bubbles', further limiting the chance of receiving interesting and relevant information from outside of the immediate social circle. In this thesis, Twitter is used as a platform for researching methods for deriving "interestingness" through popularity as given by the mechanism of *retweeting*, which allows information to be propagated further between users on Twitter's social graph. Retweet behaviours are studied, and features; such as those surrounding Tweet audience, information redundancy, and propagation depth through path-length, are uncovered to help relate retweet action to the underlying social graph and the communities it represents. This culminates in research into a methodology for assigning scores to Tweets based on their 'quality', which is validated and shown to perform well in various situations.

Contents

List of Publications

Some of the work produced towards this thesis has also been published separately as follows.

- [58] - W. Webberley, S. M. Allen, R. M. Whitaker. *Inferring the Interesting Tweets in Your Network*, in *Workshop on Analyzing Social Media for the Benefit of Society (SOCIETY 2.0), 3rd International Conference on Social Computing and its Applications (SCA)*, Karlsruhe, Germany. *IEEE 2013*

- [57] - W. Webberley, S. Allen, R. Whitaker. *Retweeting: A Study of Message-Forwarding in Twitter*, in *Workshop on Mobile and Online Social Networks (MOSN'11), 5th International Conference on Network and System Security (NSS)*, Milan, Italy. *IEEE 2011*

List of Figures

List of Tables

List of Acronyms

MT Modified Tweet

MTW Mechanical Turk Worker

OSN Online Social Network

RT Retweet

URL Uniform Resource Locator

Glossary

Audience

The set of users that receive a given Tweet, either directly or as the result of retweets of that Tweet.

Author

A user that has written a Tweet. The original author of Tweet t is denoted as $t.\text{author}_O$.

Follower

A user that follows another user on Twitter. Users who follow a particular user will receive all of the latter's Tweets and retweets to their home timeline. A user can elect to follow another user. The subset of users that are followers of user u is denoted as $N^+(u)$.

Friend

The inverse of follower. User *x* is a friend to user *y* if *y* follows *x*. The subset of users that are friends of user u is denoted as $N^-(u)$.

Local Graph

The local graph of a user, u, is the subgraph of the full social graph representing the users and edges existing within n hops of u. In the scope of this thesis, local graphs of users are limited to the subgraph generated using $n = 2$.

Path-length

The penetration of a Tweet - i.e. the number of times a Tweet is retweeted down one chain. The distance between the leaf user and root user of the chain indicates the number of hops the Tweet has taken from its author down this chain. For more information, see Definition 3.2.

Retweet

n. - A replica of a Tweet, which has been forwarded on by a user (who is not the Tweet's original author) to their own followers. The set of retweets of a given Tweet, t, is denoted as $RT(t)$.

v. - The act of replicating a Tweet. A user who finds a Tweet interesting may retweet it so that it gains more exposure through an increase in the audience size (see above).

Retweet Group

A set containing all of the retweets of a particular Tweet, t, and t itself. The retweet group of t is denoted by $RG(t)$, and has a minimum cardinality of 1 in cases where t has not been retweeted. For more information, see Definition 3.1.

Retweet Count

The number of times a particular Tweet has been retweeted. The retweet count of Tweet t is denoted as $t.\text{count}_R$ and has a minimum value of 0 in cases where t has not been retweeted.

Social Graph

The representation of users and the links illustrating relationships between them in social networks.

Timeline

A set of Tweets displayed in reverse-chronological order. A **user** timeline consists of that user's Tweets and retweets created by the user. A user's **home** timeline consists of that user's Tweets and retweets, the Tweets of each friend of the user, and retweets created by friends of the user. The home timeline contains all of the information that the user directly receives.

Tweet

n. - A piece of information in Twitter; a piece of text, no more than 140 characters long, which is written by a user. When sent, the Tweet is pre-pended to its author's user timeline and also to the home timelines of each of the followers of the Tweet's author. A Tweet is denoted as t.
v. - The act of writing and sending a Tweet.

Note - A Tweet, in the context of Twitter, is treated as a proper noun and as such has its first letter capitalised[1].

User

An account on Twitter. Each user (usually representing a real-life person or organisation) can Tweet, retweet, follow other users and be followed by other users. In this thesis, the terms *user* and *person* are occasionally used interchangeably.

[1] https://twitter.com/logo

Chapter 1

Introduction

Online social networks have exploded into the lives of millions of people worldwide over the last decade, and their use has facilitated the interconnection of the world in ways never before perceived possible.

These social networks have many characteristics that are also exhibited by 'real'-world social networks. Although most such platforms provide a different service to collaboratively satisfy an array of different use-cases, they tend to all be based around the idea of 'friendships' (i.e. links between the user nodes in the social graph) and the sharing of information amongst friends.

Social networks like these have been available for around ten years now (with MySpace[1] launching in 2003 and Bebo[2] in 2005), but it wasn't really until the worldwide launch of Facebook[3] in 2006 that social networks became the staple, ubiquitous norm that they are today. More recently, we have seen the introductions of Google's social network grown from its Buzz service, Google Plus[4], Pinterest[5], App.net[6], and many more. They make up a large part of and contribute heavily towards the ideas behind Web 2.0, which describes the web as being primarily formed from user-generated content and encourages the sharing of such content.

[1]http://myspace.com
[2]http://bebo.com
[3]http://facebook.com
[4]http://plus.google.com
[5]http://pinterest.com
[6]http://app.net

Another component that helped in the dawn of Web 2.0 was the rise of *blogging*. A blog ('web-log') is a time-based series of posts consisting of continuous pieces of text, photos, or other media, and is generally contributed to by a single author. Blogs are often based around one or a set of topics and are usually public - meaning that they are written with the intention of being read by others. Despite this, they are often a way in which the author can look back at their history of posts, acting more as a diary recording snapshots of the author's life. Various blogging services exist on the web today, such as Medium[7], Wordpress[8], and Tumblr[9].

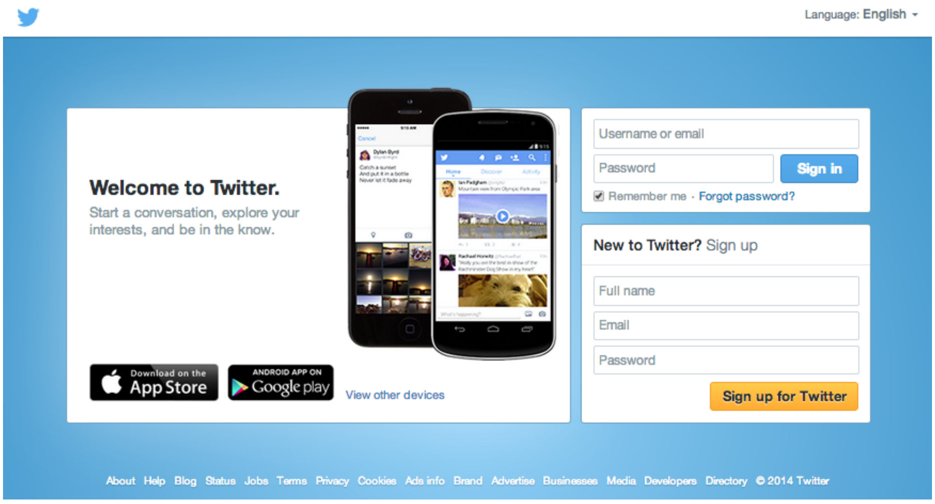
1.1 Twitter as a Social Network

Twitter[10] is an online social network, which launched in the summer of 2006 [34]. Since then, it has rapidly gained in popularity amongst several different user groups - teens and young people, casual users, celebrities, reporters, and so on - and within eight months had around 94,000 registered users [32]. Whilst the design of the site and its apps has changed significantly since its launch (Figures 1.1(a) and 1.1(b)), its function has remained mostly constant. Twitter has never been a direct competitor with Facebook, as users tend to use the two sites concurrently for different purposes: whilst Facebook's focus is on providing many services at once (such as photo-sharing, commenting/endorsing of information, messaging, pages for businesses, groups, events, etc.), Twitter's is more on simplicity.

More specifically than just being an online social network, Twitter is a microblogging website. Whilst a blog, as mentioned, typically contains long posts, Twitter only allows its users to post short pieces of text, up to 140 characters in length [34, 30], called 'Tweets'. Thus, Twitter is a hybrid social network and blogging service and whilst

[7]http://medium.com
[8]http://wordpress.com
[9]http://tumblr.com
[10]http://twitter.com

(a) Twitter's landing page in 2006.



(b) Twitter's landing page in 2014.

Figure 1.1: Twitter's 2006 homepage (from http://web.archive.org) compared to its 2014 homepage.

each Tweet may only realistically be able to hold a couple of sentences, this system facilitates quick, timely, and 'real-time' *live* information-sharing amongst its millions of users [65]. Its idea is that short pieces of news will 'travel' faster and will be seen by more people more quickly than traditional news stories.

Although Tweets are limited to 140 characters in length, the inclusion of URLs is allowed. This enables further extension of Tweets through external websites, and supports the inclusion of links to images and videos. Twitter has encouraged this use-case by providing 'share' buttons for developers to embed in websites, and direct support for photo and video applications, such as TwitPic[11] and Vine[12].

Its simplicity has also helped its growth into the mobile domain, in which smartphone users are able to very quickly post updates about their lives, a piece of information they want to share, or a photo or video, and be able to post it *as it happens* directly from the news source or geographical location [13]. This has been especially useful in emergency situations worldwide, including the Haiti earthquake in 2010 [42], and 2011's Egyptian protests [59] and Thai flood [33]. Indeed, Sakaki et al. [50] used Twitter to build an earthquake-reporting system for Japan that outperforms the Japan Meteorological Agency in terms of its promptness of notification.

Use of Twitter is based around 'timelines' of Tweets, to which new Tweets are prepended as they are posted by users. The *home* timeline is the default view, in which Tweets from all of a person's subscribed-to users are placed. Timelines of an individual user contain only Tweets from that user, and are known as a 'user' timeline. Customisation of timelines is also possible through the use of Twitter lists, in which different users can be placed to categorise streams of Tweets from different sets of users.

1.2 The Social Graph and Information Flow

The structure of Twitter lies within the users and their connectivity within its social graph. However, unlike Facebook, whose social structure is made up of bi-directional 'friendships' between users, Twitter's primary social graph is made up more of mono-directional links between its users [19]. A person using Twitter can elect to *follow*

[11] http://twitpic.com

[12] http://vine.com

another user, which subscribes the person to receive all of that user's Tweets to their home timeline. The set of users that follow a person are known as that person's *followers*, and the set of users that the person follows are the person's *friends*. Therefore, if two users both mutually follow each other, then the link between them is bi-directional.

Whilst bi-directional links are common amongst communities of similar interests, friends, colleagues, and so on, mono-directional links are found more in situations in which less-influential users follow more-influential users, such as celebrities.

1.3 The Problem

A person who follows a set of other users can generally be assumed to find them to produce more interesting information than those that the person does not follow. However, despite that, not *all* information produced by a followed user is likely to be interesting to the person, and yet *all* information produced by a Twitter friend will be received onto the home timeline.

Noise is a common problem in Twitter, and is the uninteresting information one might receive that conveys little interest. It is likely that most of the information received on Twitter *is* uninteresting [4], and this makes it very hard to distinguish the interesting from the uninteresting.

Since people tend to use Twitter most in short and sporadic moments, looking for a quick news or information fix, they do not have time to filter out noisy Tweets. Thus, the presence of noise can dampen the user's experience, making it much more difficult to find the interesting information.

In addition, Twitter users typically exist within an information 'bubble'. This is similar to the notion of the Google search bubble, in which the search engine uses previous results and search terms to only return information to a user based on what *it thinks* the user would find the most interesting and useful. This results in the users not knowing which information exists beyond the confines of their bubble, and if they do not know it exists, they cannot know if it is of interest to them. Similarly, a Twitter user cannot follow all of the users he/she may find interesting, since he/she will not *know* of all the interesting users existing on the social graph.

Although not directly answered in this research, the key theme and motivation behind the work it involves is;

How can users be exposed to *interesting* and *relevant* information, but without explicit subscription or search?

The remaining chapters aim to go some way to answering this question through a focus on understanding information propagation, and how combining this with knowledge of the social structure of Twitter can assist towards solving the problem of identifying interesting and relevant information and determining it from the noise surrounding it. Whilst other work in the area, discussed further in the Background chapter, has also researched into the notions of relevance and interest in online social networks, and Twitter in particular, they do not address the problem in the same way or attempt to validate findings so thoroughly.

The central thesis addressed by this research is that the interestingness of Tweets in Twitter can be non-semantically inferred through consideration of the underlying social graph model and the use of humans as proxies to the perception of interesting information. In this work, the following research questions help illustrate the steps towards realising this thesis:

- **RQ1** *Does Tweet popularity, measured in terms of retweets,* ***define*** *interesting information?*
- **RQ2** *Can Tweet popularity, measured in terms of retweets,* ***be an indicator*** *of interesting information?*
- **RQ3** *Is the arrangement of Twitter's social graph an important factor in retweet propagation, and thus perceived popularity?*

- **RQ4** *Can Tweet interestingness be inferred* ***non-semantically****?*

In answering these questions, the main contributions made in this research work are outlined below.

- Tweet analysis allows for providing definitions for key concepts, such as retweet groups and propagation path-length, to form foundations for further definitions and work towards substantiating the thesis.

- Social graph analysis shows how the underlying edges dictate information flow through Twitter, and that altering the network structure has a clear effect on the propagation characteristics of Tweets within it.

- A mathematical 'definition' of global interestingness is provided, given as a score for the purposes of quantifying and ranking Tweets by this metric. The score is an estimation based from a Tweet's predicted popularity.

- Rigorous validation tests are conducted to demonstrate the score's ability to identify useful Tweets from noise and rank information by interestingness. User tests involving Tweets 'local' to them on the social graph show that there is scope for the scores to also be applied locally as well as globally (for addressing information relevance).

1.4 Thesis Structure

The chapters containing the remainder of this thesis are laid out as follows.

- **Background and Research Domain** - A chapter providing an introduction to key concepts and the ideas behind the main research. Also included is a review of relevant literature across the range of topics addressed in the thesis.

- **Understanding the Behaviour of Retweeting on Twitter** - Initial research into the problem areas identified, including studies into the propagation characteristics exhibited by Twitter and its useful and interesting properties. This is in order to provide an understanding of the mechanics of retweets.

- **Analysis of Twitter's Social Structure** - An in-depth analysis into the layout of Twitter's social graph and the ways in which it helps govern the potential spread of retweets. From this graph analysis, an initial method for inferring Tweet interestingness emerges.

- **Inferring Interestingness** - Improvements to the inference methodology are introduced, and deeper validations and analyses are conducted into the relative strengths of the new method. Different types of assessments are conducted in order to investigate the performance on a general scale as well as moving towards addressing information relevance.

- **Assessment and Conclusions** - The thesis ends with a general analysis and conclusion, and a discussion of potential future work in this area and leading on from this research. The contributions are summarised and the research questions answered more formally.

Chapter 2

Background and Research Domain

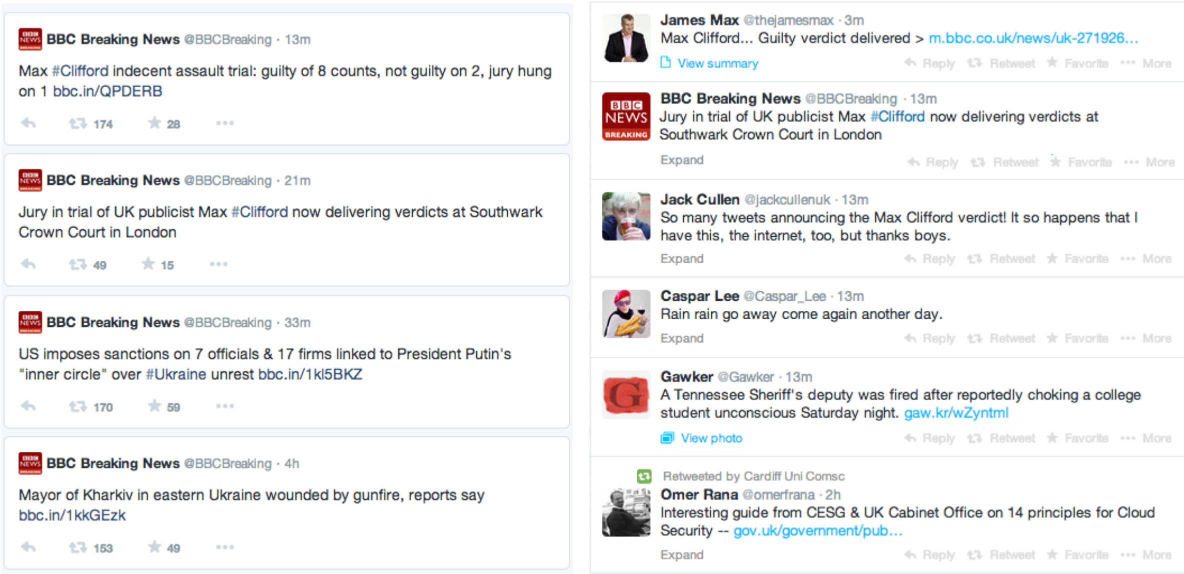
One of the most widely-used features of Twitter is its inbuilt function for easily facilitating the spread of information through its social structure. This phenomenon is the basis for much of the research in this thesis and, when combined with the characteristics of Twitter's user graph, has many interesting attributes and behaviours associated with it.

2.1 Key Concepts

Although the main important concepts are described briefly in the Glossary of this thesis, in order to more clearly understand the research and components of Twitter they are explained in more depth in this section.

Tweet

A Tweet is a singular message on Twitter that cannot exceed 140 characters in length. It is written and sent by a user, who is the Tweet's author. Figure 2.5 shows a Tweet written by a user, which is pre-pended to the author user's user timeline and to all of the author's followers' home timelines. Tweets can be retweeted, as explained later in this chapter, in order to create a retweet. The term 'tweet' can also be a verb defining the act of sending a Tweet.



(a) The user timeline for user `@BBCBreaking`.

(b) A home timeline.

Figure 2.1: Examples of user and home timelines.

User Timeline

The user timeline of user u is a set containing the Tweets written by u and any retweets made by u and is ordered by time with the most recent Tweets at the top. Figure 2.1(a) illustrates this through the example of BBC Breaking News' user timeline. Since user timelines also contain retweets, not all Tweets on u's user timeline are necessarily authored by u.

Home Timeline

The home timeline of user u is a set containing the Tweets written and retweeted by u as well as all of the Tweets and retweets made by his/her friends, where a friend of u is another user that u follows. Figure 2.1(b) shows an example of a home timeline. As such, a home timeline is the union of the user timelines of u and all of u's friends, again with the most recent Tweets nearer the top.

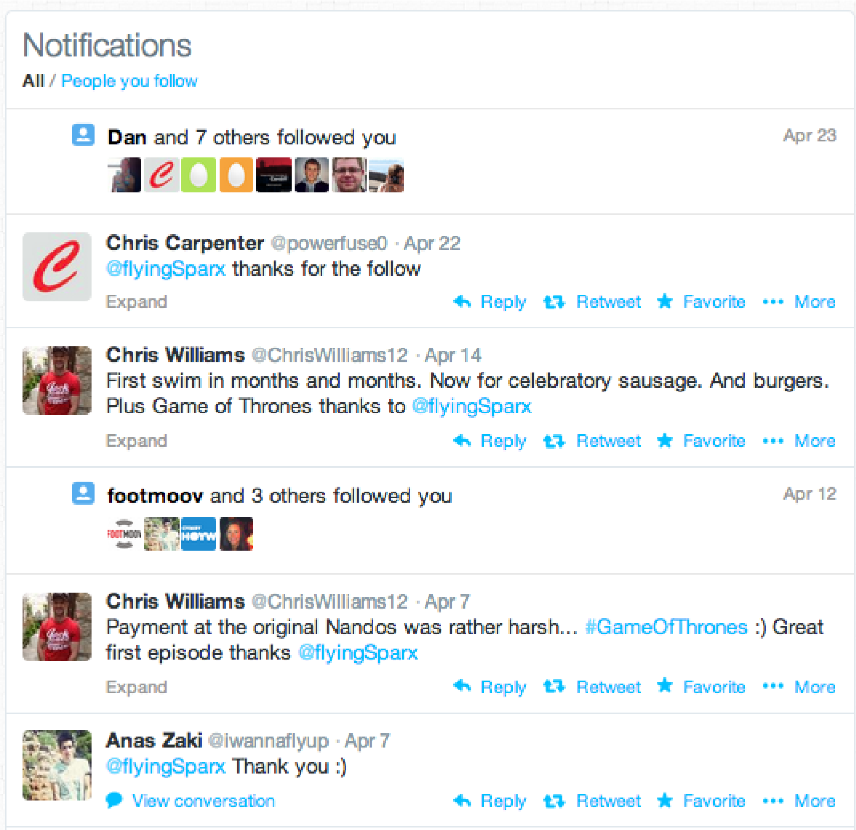


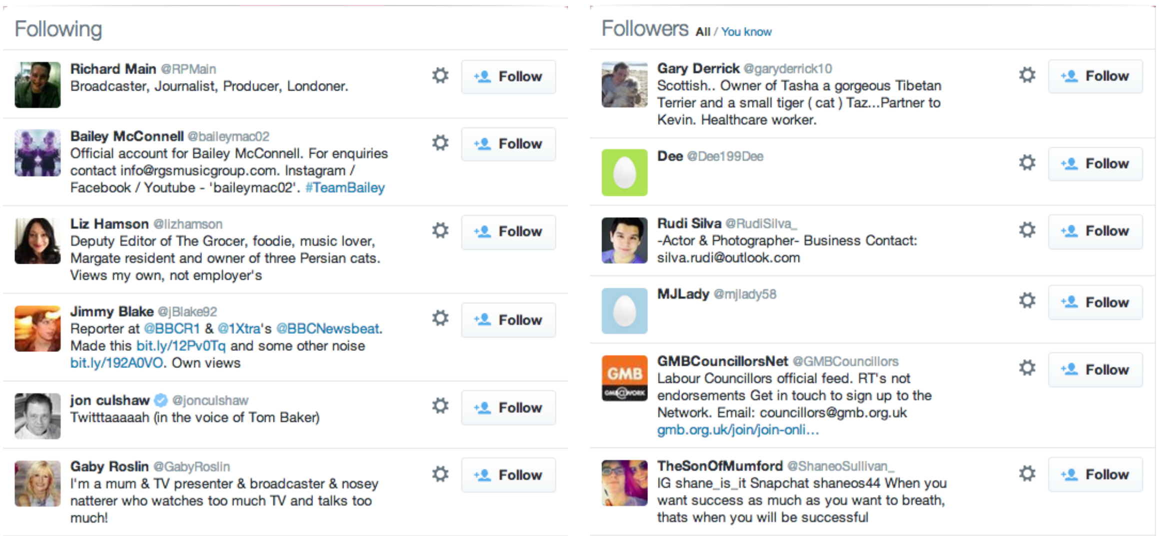
Figure 2.2: A notifications page.

Notifications Page

The notifications page for a user u is a page that only u can access, and which lists events involving him/her. For example, this may contain new Tweet mentions (Tweets that contain u's username) and new followers, as shown in Figure 2.2.

Friends List

The friends list of user u is the set of users that u follows ordered with the most recently-made friendships nearer the top. Therefore, the first user u followed is at the bottom of his/her friends list. Figure 2.3(a) shows an example of a friends (or 'following') list.



(a) A friends or 'following' list. (b) A followers list.

Figure 2.3: Examples of friends and followers lists.

Followers List

The followers list of user u is the set of users that follow u. Again, this is ordered with the most recent follower at the top. An example of a followers list is given by Figure 2.3(b).

2.2 Domain Introduction and Literature Survey

This section provides an understanding of retweeting in Twitter and its effects on the users and the followships between them, which represent the underlying social graph of Twitter, and includes a survey of some of the most relevant works of the area.

2.2.1 Information Propagation through Retweeting

The function of propagation in Twitter is known as *retweeting*, and is carried out by the Twitter users themselves. When a user views a Tweet that they believe to be particularly interesting, and believe it to also be interesting to his/her followers, then he/she can elect to retweet it, and thus pass it further through the social graph to that user's followers also. A Tweet that has been retweeted is known as a *retweet*, and it is clear that a Tweet which is retweeted will be made available to significantly more users than a Tweet that isn't retweeted [57, 35]. Since Twitter's social graph is decentralised and retweeting occurs between individual groups of users, its properties are similar to information dissemination in other types of decentralised graphs, such as content-forwarding in opportunistic networking [3].

A retweet can be carried out in one of two ways: either through the use of Twitter's native retweet 'button', or manually. The button is displayed along with each Tweet (see Figure 2.5) in a timeline which, when clicked, immediately creates a new retweet containing the verbatim content of the original Tweet and automatically sends it on to the retweeting user's followers. The user who created the original Tweet is credited as the author on the recipients' timelines, with an indication of who carried out the retweet itself. Thus, users who follow the retweeter will see a Tweet appear in their home timeline from someone that they may not directly follow. Figure 2.4 illustrates an example; the user receiving the depicted Tweet does not follow the original author, @aldakaida, but *does* follow @DTW_Holidays, who was responsible for carrying out the retweet.

Figure 2.4: A retweeted Tweet.

The manual approach involves physically copying the content of the Tweet to be retweeted and pasting it into a new Tweet, usually with the text 'RT @<username>:' prepended, where RT stands for **ret**weet and <username> is the username of the author of the original Tweet. This method allows for annotating the original content of the Tweet (for example, to provide an opinion on the Tweet contents), producing a *modi-*

fied Tweet, which can sometimes be pre-pended with `MT` rather than `RT`.

Historically, this manual approach was informally community-driven by the users of Twitter and was the only method available for carrying out retweets. However, in 2009, Twitter realised the popularity of this user convention and introduced the retweet button[1] in order to assist in this trend, and through which Tweets could be retweeted much more quickly and accessibly. The button is implemented on Twitter's website and mobile device applications, yet even today the original and manual retweet approach remains popular amongst many of Twitter's users and communities.

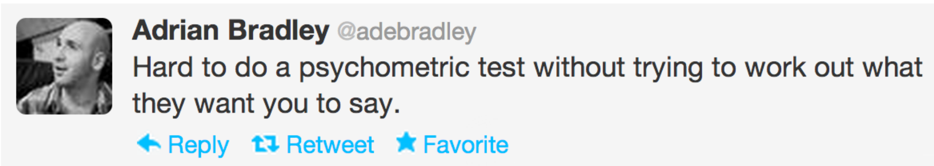


Figure 2.5: The retweet 'button' in context.

Each Tweet has a retweet count associated with it, which is the raw representation of the number of times that the Tweet has been retweeted using the retweet button method. Since the manual retweet technique remains more community-driven, there is no official way to include these as part of the retweet count of the original Tweet. However, since the manual method is typically only used with the aim to annotate or modify the Tweet in some way, the resultant 'retweet' is no longer a real representation of the content of the original Tweet, and so should not be counted as such.

It should be noted that Twitter users may choose to make their account 'protected'. A person who has a protected account will still have a publicly-visible profile (displaying a name, username, bio, and so on), but their Tweets and other information (such as the followers and friends lists) are hidden from users that aren't followers of the person. Potential followers of a protected account must *request* a followship, which can then be accepted or rejected by the protected account holder. Since Tweets from a protected account are only visible to approved followers, the retweet button is unavailable for

[1]https://blog.twitter.com/2009/project-retweet-phase-one

them to disseminate the Tweet any further than the author's immediate local follower network. However, since the manual retweet method does not rely on the button and isn't governed by Twitter, a protected account's Tweets can still be retweeted in this way.

As Facebook supports the endorsement of information found on its site by inviting users to 'like' a piece of content, retweeting is effectively a *vote* or endorsement for a Tweet on Twitter. In both cases, the number of likes and number of retweets is available to the platforms' respective users (Figure 2.9), and so this provides some insight into the *popularity* of the information. Whilst Facebook 'likes' are immediately visible to users in feeds, the retweet count becomes visible once a user clicks a Tweet to expand its metadata.

However, 'likes' and retweets may not have the same meaning. Facebook users 'like' a post or a photo if they agree or enjoy the content, whereas a Twitter user retweeting a Tweet is an indication of its relevance.

Some Twitter users declare that their 'retweets are not endorsements' in their profile's bio. This particular behaviour is largely associated with journalists who use retweets to highlight potentially controversial Tweets and to ignite discussion over certain topics. This declaration is unwelcomed by some users, who then argue the point by asking what retweets *are* meant to imply[2] and that a user's bio is not immediately available for recipients to view in Tweet streams. Generally, this is not very widespread, and even a user retweeting a Tweet with the aim of it not being an endorsement implies that particular user's interest in the Tweet.

[2]http://www.poynter.org/latest-news/media-lab/social-media/152448/the-problem-with-retweets-how-journalists-can-solve-it/

2.2.2 Retweets and the Social Graph

The social graph of Twitter can be represented, like other online social networks, by edges between users, partially emulating real-life social interactions between humans. The growth of social media has encouraged more dense communication between users all over the world, who would not previously be able to be in direct contact with one another in this way.

Stanley Milgram's "six degrees of separation" [38] experiments are highly relevant to and useful for OSN research today. The results of the experiments demonstrated that people are usually no more than six hops away from each other on the world's social graph, yet this value was found to be an overestimate when it comes to the analysis of the structure of OSNs by Backstrom et al. [7], who found that the average 'distance' observed in Facebook's entire 721 million-node graph in 2011 was only around 4.7 hops. This implies that denser links between users and larger communities that apparently manifest themselves in OSNs create a smaller 'world' than that experienced in reality.

In each of Milgram's experiments participants passed a message to one another, at each stage only passing to other people that they actually *know*, in the hope of it reaching a single intended recipient. This meant that people could use acquaintances in other geographic locations to transfer the message from community to community. Twitter supports a similar propagation mechanism in the fact that retweets can themselves be retweeted; this is a focus of some of the earlier research in this thesis.

This behaviour provides further penetrative 'depth' of the information through the social network away from the source user in addition to the spread 'width' made by the initial retweets. Although retweeting is not carried out with the aim of information reaching any particular final user (or set of users), as with Milgram's experiment, this phenomenon allows retweets to 'travel' between 'online communities' of users.

As with real-life social networks, communities of users in OSNs are also a common feature [55]. Social communities are identified by clusters of people sharing dense links with each other, and can vary in size, location, and geographic spread. They often grow over time and are formed between people living near to each other or between groups of users with shared interests.

In Twitter, these communities are typically small to begin with and are based on a topic of interest or around a more influential user. As Tweets are produced from within the community, further links are made to interconnect the community's users, producing a growing 'swarm' of interest around the initial topic or user [32]. As further users begin associating themselves with this community, its audience becomes widespread and the community grows. This concept is discussed in greater length by Java et al. [32], who described them as compact groups of users connected by dense follower links after further experimentation.

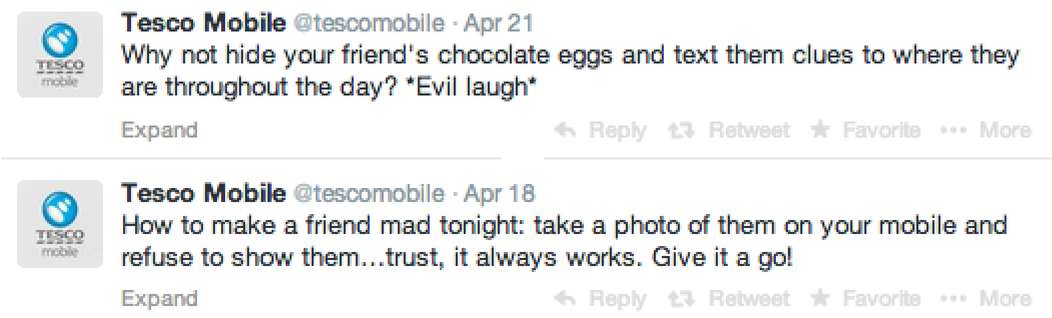


Figure 2.6: Phatic Tweets from Tesco Mobile's official Twitter account.

Some users try and build communities through *phatic* communication, in which the Tweets being broadcast are 'non-dialogic' or 'non-informational', as defined by Miller [39]. This is often done with the aim of providing a relaxed environment, and could be useful for cases similar to an internal Twitter account for an organisation's employees. Additionally, some commercial organisations may use phatic language in order to engage more with the community for the purposes of appearing more approachable, such as Tweets from Tesco Mobile's official Twitter account, as demonstrated by Figure 2.6.

In more dense communities, Tweets can be made available to many users immediately

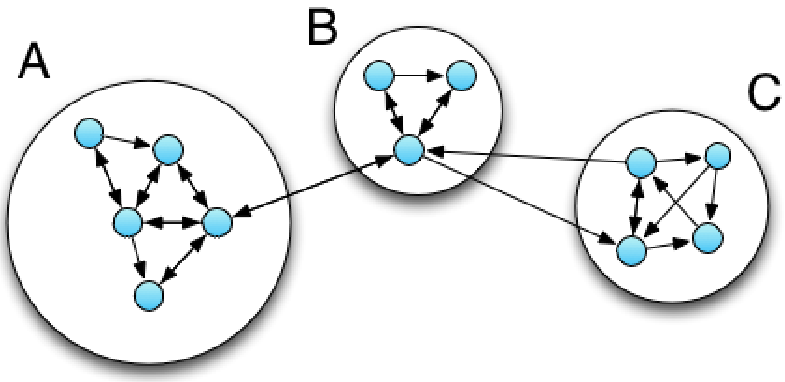


Figure 2.7: A hypothetical group of user communities.

after they are published, since many of the links between users are shared. This means that any retweets that occur within communities are likely to have a lot of *redundancy*, in that many of the retweets will be sent to users who have already seen the original Tweet. Twitter prevents this information duplication by not displaying the retweets of Tweets that have already appeared on a user's timeline. However, this does increase the chance of the Tweet being propagated to users external to the community.

Retweets amongst users within a community are likely to be common, due to the shared-interest nature of communities, and some users can provide 'bridges' by being active in more than one group. In these cases, Tweets can be passed between the communities through retweets by the bridging user. For example, Figure 2.7 illustrates three hypothetical communities 'bridged' by one user in community B. If this user was to retweet a Tweet from communities A or C, then it is clear how the Tweet could be propagated from one group of users to another. If there are many users sharing communities then there are many more avenues available for dissemination through the graph, causing a high level of information throughput. If there are fewer bridges, then there is more of a bottleneck between the communities, hindering the information spread.

Java et al. [32] also find that communities can be formed from different types of people, such as those who Tweet frequently and have many followers, and those who contribute very little and have few followers. Those with many followers and many friends receive lots of information and have the potential to spread information further than those with fewer inward and outward edges. Studies into the behaviour of different types of user in Twitter is done more thoroughly by Krishnamurthy et al. [34], who define ‘broadcasters’ (users with many followers and few friends) and ‘miscreants’ (users with few followers but many friends) and their roles in information propagation. The authors describe broadcasters as users who post many Tweets, which are then received by large numbers of users, and miscreants as those who receive lots of information but are unable to achieve a large Tweet audience. It is therefore assumed that broadcasters are likely to be retweeted more than many other types of user.

Users that retweet the interesting information from a source user to others, who do not follow the source user and so would not naturally receive the information, are effectively acting as information *filters*. By not following the source user, a person might still receive the interesting information through these filters, but will not receive any of the ‘noise’. Thus retweeting means that friends of a user become useful filters of information for users further ‘downstream’ and retweeted information can be said to have a higher *credibility* than Tweets that aren’t retweeted [13].

2.2.3 User Influence

Just as there are different types of user *behaviours* on Twitter, there are also users of different *influence* levels [49]. Much research has gone into user influence, including on how this might be detected [62], and influential users are generally found to be those who have a greater impact on Twitter’s social network [8] and who usually have significantly more followers than an average user. Influential users tend to have a high persuasion over other users, relating *influtentials* in Twitter to those who are also influential in the real world as part of traditional communication theory [15], and therefore many Twitter influentials are the accounts belonging to real-world celebrities.

As with real-world celebrities, Twitter influentials are those with many 'influenced' followers, or fans, which are the users who have the strongest agreeable opinions of the influential. As a result, an influential user has a greater number of followers who are interested in the information produced by the user, and is therefore more likely to receive more retweets than less influential users.

Although influence level is partly derived from the follower count of the user, it should be noted that a user with high in-degree on the social graph[3] does not necessarily imply a high level of influence. An 'active' audience of users who reply, retweet, and interact are more indicative of an influential user [10]. This is especially true since a user can gain more followers through campaigns such as '#teamfollowback'[4] or by following 'out of politeness', in which a user will follow another user back as an act of politeness, but these users tend to have *both* high in- and out-degree and invoke less interactivity amongst their followers, which are not necessarily characteristics of an influential user [15].

Klout[5] is a web service that attempts to review a user's social media influence by assigning users a Klout Score. Their website declares that this score, which ranges from 0 to a maximum of 100 and whose generation algorithm is kept private and unpublished [19], is determined from a variety of over 400 'signals' taken from eight different social media platforms. These signals are derived from various attributes including the volume of information shared, the reaction to the shared information, and the relative scores of the users who interact with the information. They *also* take interactivity between users as one of the primary indicators [5]. Additionally, the service indicates the topics a user is influential about, with the general idea being for organisations to check up on which users are influential for marketing purposes, but also to highlight the users that should be replied-to at a higher priority.

[3]In-degree: many followers.

[4]Users associate themselves with #teamfollowback to imply they will return all followships.

[5]http://klout.com

2.2.4 Twitter as an Information-Retrieval System

At a high level, Twitter can be considered a type of information-retrieval system, which people can utilise to produce and consume information when required. In traditional information-retrieval systems, such as search engines and library systems, keywords and search terms are common ways for describing the type of information the user would like to receive. The system would then search a database or archive for what it believes is relevant information, *based* on these 'retrieval parameters', and return results to the user ordered usually by the estimated relevance of the articles [6].

Information quality is also reliant on the expected reading effort of the returned documents. The character precision-recall metric was introduced by Arvola et al. [6] by way of demonstrating the tolerance-to-irrelevance ratio. The general mechanism for this ratio is centred around users reading a document passage; the point at which the ratio is reached is when the user stops reading the passage and moves to the next whole document since they assume the rest of the current document will also be irrelevant to them. Therefore, the more effective the information retrieval system is in displaying high-quality information, the lower the chance that this ratio is reached by the user. This is comparable to the event in which a Twitter user viewing Tweets from someone they are following gets to the point where he or she reaches this ratio (i.e. is beginning to get bored or find the Tweets irrelevant) and decides to unfollow the friend. Similarly, the more effective the user is when selecting people to follow in the hope of receiving interesting information, the less likely it is that the user will remove these friends.

Whilst Twitter does not support the use of keyword searching for its primary information delivery method, it does lend its users some control over the type of information they wish to receive. As mentioned previously, users receive all of the Tweets from everyone that they follow onto their home timelines. Thus, by selecting users to follow, a person is effectively describing and implicitly indicating the type of information he/she would like to receive, and by editing their friends list (either by adding new followers or pruning existing ones) he/she can alter this indication. In addition, Twitter moves towards an information-*recommendation* system, in that the friends of a user can endorse and imply a Tweet's quality by retweeting it.

Despite this control, it is still unlikely that users will achieve a perfect Twitter experience due to the presence of *noise* [4]. As discussed in the Introduction, this problem stems from that although a person follows users they consider to be interesting, it is often the case that not *all* information produced by interesting users will be interesting itself. For example, a user may follow a news source in order to receive breaking news but is not interested in viewing the information about politics or celebrity updates that the news source also Tweets about.

Twitter has gone some way to acknowledge these issues, especially with the introduction of the 'mute' feature[6], which allows users to mute their friends. However, the result of this, with respect to the home timeline, is equal to unfollowing a friend and so does not fully solve the problem.

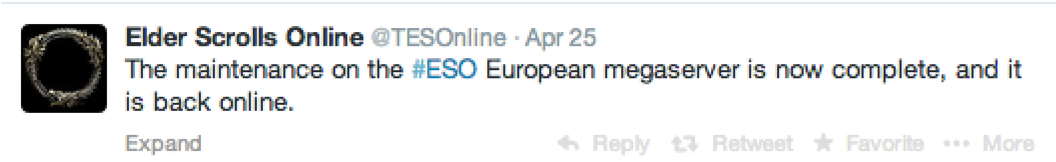


Figure 2.8: Example of a Tweet from a 'dynamic' Twitter account.

Twitter's information-retrieval characteristics are also very valuable in highly dynamic scenarios, where status updates regarding rapidly-changing circumstances are very useful. There are some Twitter users that might use the service in this way very often, for example a user who may post updates on server availability for an online game, as shown in Figure 2.8, or a user who broadcasts Tweets on the state of transport links, such as the London Tube.

In addition to 'permanent' dynamic behaviour, environments of uprising and political tensions, such as the 'Green Revolution' during the 2009 Iranian elections [11], clearly demonstrate Twitter's important role in the live updating of information and news in

[6]http://blog.twitter.com/2014/another-way-to-edit-your-twitter-experience-with-mute

difficult situations. The examples nearer the start of the Introduction chapter also help illustrate this.

2.2.5 Interesting and 'Interestingness'

A person will find a piece of information interesting if it is relevant to them and conveys some amount of 'affective stimulation'. In Twitter, a person will attempt to arrange their local network, by following others, in order to improve the likelihood of receiving interesting Tweets. However, in most cases, a Twitter user electing to follow another cannot predict precisely what the new friend will Tweet about in the future. The user has some *expectation* of the type of information they are likely to receive based on the previous Tweets of the new friend, which is generally the main cue the user can use to base the follow decision on.

Part of the follow decision is based on the notion of relevance judgement, which is an idea discussed at more length by Xu [60] and is partly made up of the goal of achieving affective stimulation through *hedonic* searching as opposed to the use of *epistemic* searching.

An epistemic information search is one with the purpose of finding out information on a particular topic (or set of) to satisfy a *desire for knowledge* [60], yet without an actual aim to solve any particular problem.

An example of this type of search is a 'crawl' through Wikipedia, in which a searcher may start at one particular page of interest and then follow links within that page to other related pages of interest that stem away from the source topic. In this case, the search 'parameter' is simply the name or title of the article the searcher wants to view. As mentioned previously, a followship between users is effectively a search parameter in Twitter, since the following user has elected to follow the new friend to receive information from him/her. It is clear that this type of 'searching' cannot be epistemic as the following user cannot know exactly the *type* of information they are going to receive.

Hedonic searching is similar to epistemic searching in that it is also not carried out with the aim to solve an immediate problem, but is different in that it is done to find fun or ‘affective stimulation’ [60]. A person can be said to be affectively stimulated if he/she views a piece of information that has some effect on the person, such as something that conveys emotion, something that is of particular interest to the person, or something that is capable of provoking some further thought.

With hedonic searching, users are not aware of the information that they are going to receive prior to searching and thus cannot really predict any level of affective stimulation. This aligns more with Twitter usage, since users receive information that they cannot accurately foresee. Any Tweets received that do provide interesting information can convey affective stimulation to the user. This is the type of information that becomes harder to identify amongst lots of noise, yet is also the type of information a user is more likely to retweet.

Interesting content is not simply that which someone enjoys reading, since an article that conveys anger or frustration is also stimulating, and thus this concept describes information that a user is curious or concerned about.

Interestingness is used as a concept in other domains also, including in pattern and data mining. Geng and Hamilton [23] survey nine approaches for measuring interestingness in data mining. In this paper, the authors assess metrics such as peculiarity, surprisingness, generality, and diversity for deducing levels of interestingness. The approaches tend to use semantic methods for measuring in different categories, such as objectively, subjectively, and in mearsuring for the purposes of classification.

Mitra et al. [41] survey approaches for data mining using soft computing, some of which use features such as validity, novelty, and understandability to describe interestingness. The authors also discuss the use of the metric as a threshold for reporting useful information. *Relative* interestingness is a term used by Hussain et al. [31] for finding the best (most interesting) rules to make use of mined data. The methods use common sense knowledge to mine rules that contradict a user's knowledge, on the basis that this form of contradiction is bound to be interesting to the user.

Tan et al. [54] present a method for finding the best measure by which to reduce mined data. The method highlights the most useful and interesting properties to consider when reducing data across multiple domains, such as in support-based pruning and table standardisation. Similarly, Silberschatz and Tuzhilin [51] measures interestingness of mined patterns based on whether the data is unexpected or actionable to a user, where information is interesting to a particular user when contrasting with that user's beliefs. Information affectiveness has a clear link with interestingness in user decisions on whether to accept or pay attention to incoming information. Hidi and Baird [27] argue how the study of affect on a user is imperative to understanding what is interesting to the user, and that this can be sourced from stories (containing features such as suspense), news articles, and events. The authors report on how interest driven by affect increases the cognitive application to the discourse (or input information), and increases recall and learning capacity about the semantic topics it contains.

It is clear that the concept of 'interestingness' is distinct from 'interesting', in that the former is an attribute of information and the latter is in the possession of a user perceiving information. The notion of affective stimulation and its links to information interest and relevance has been explored, yet due to the subjective nature of information perception, the property of *interestingness* cannot be defined in the same way. The research discussed above makes clear how information can have the property of interestingness if it contradicts general belief, is peculiar, surprising, novel, useful, and so on. Generally, therefore, interestingness of information applies to that which 'differs from the norm' and is representative of the cumulative interest of everyone who evaluates a piece of information.

Stemming from the generic research into interestingness, it is used in this thesis as a term for describing Tweets that are 'not noise', and contain some amount of generally (or 'globally') interesting content that is different to *what is expected*. This theme is key to the interestingness metrics discussed later in this research, and the notion of a threshold, as in [41], is useful for *determining* the Tweets that might have the attribute of interestingness.

This global interestingness can be used as a measure of information *quality*, as it denotes information that stands out from what is mundane. Information-retrieval systems typically use some measure of information quality and relevance when determining which documents to return to a user and also when deciding on the order the documents should be displayed in. This 'quality' is subjective in that different systems use a variety of different algorithms for deducing quality, usually based on the level of interestingness of each of the available documents. Interestingness is useful here, since documents that many people find useful are less likely to be 'noise' and are therefore more likely to be returned. In some cases, such as with Google's Page Rank, the level of quality itself also depends on the interest of the particular user requesting the information.

Google's Page-Rank algorithm uses multiple cues to determine who the user is, their interests, past searching habits, links clicked, and so on, to return *relevant* information, which is incidentally one of the causes of the aforementioned Google search bubble. Amazon's recommendation algorithms analyse a user's past item views and purchases and cross-matches these against trends based from users who also looked or bought similar items. Amazon is then able to accurately determine the type of items a customer are interested in purchasing, and can send emails to that customer with personalised recommendations. In these cases, information quality is essentially a function of information interestingness and information relevance.

A further metric for measuring information relevance in information retrieval is the recognition heuristic. This heuristic takes advantage of a person's memory and declares that if a person is able to recognise only one of two (or more) items, then s/he is more likely to judge the recognised item to be 'greater' or more important [44, 24].

Relating this to information received on Twitter, Chorley et al. [16] found that a user recognising a Tweet's author significantly increases the chance that the user will decide to read the Tweet. Since a user must read a Tweet in order to make a decision on whether, or not, to retweet it, then the recognition heuristic transiently plays a part in a user's retweet decision also.

The authors also find that information about the Tweet itself, such as its text content and its retweet count, has much more of an effect on a user's read decision than information about the author, such as the followers count or Tweet rate. This also contributes to the declaration that information interest goes beyond the features surrounding a particular user and that user influence does not dictate interestingness of information.

Twitter itself does not attempt to deliver interesting content to its users, relying on the users themselves to implicitly 'choose' the information they want to receive. Additionally, information is always displayed in chronologically-ordered timelines, with new Tweets being continuously inserted at the top as they occur. Twitter does not try to indicate interesting Tweets on the timeline which means that the interesting information is shown at equal value alongside the 'noisy' Tweets, causing the difficulties in identifying the interesting information as has been mentioned previously. Indeed, the recent TechCrunch article from October 2013, "Twitter Quitters And The Unfiltered Feed Problem"[7] talks at more length about this particular phenomenon, and helps highlight the problem area of this work more clearly from a layperson's perspective.

The retweet count of a given Tweet is a useful metric in inferring its *popularity*. If a Tweet is retweeted 10 times, then ten people have taken the time to read that Tweet, decide it is worth sharing, and then actually retweet it [56]. This user (and the other nine retweeters) may have found the Tweet interesting, yet it should be noted that although the count can be used as a measure of popularity, as a function of the influence of the Tweet's author, the retweet count alone cannot be used as a measure of how interesting the Tweet actually is [43]. For example, it is inappropriate to say that the first Tweet

[7] http://techcrunch.com/2013/10/05/sorry-my-feed-is-full

in Figure 2.9 is so significantly more *interesting* than the second, although it is clearly more popular since Justin Bieber is an extremely influential Twitter user. Indeed, although Justin Bieber's Tweet is clearly of 'interest' to his fans, it is not necessarily the general case.

Figure 2.9: Example of Tweets with significantly different retweet counts.

Whilst the work in this thesis does not primarily aim to build an accurate retweet-predictor, this does become more important in some of the work in later chapters, since it forms part of the basis of the methods of interestingness inference as a function of many retweet *decisions*. Uysal and Croft [56] also identify the problem of 'noisy' Twitter timelines and discusses methods for predicting *popular* Tweets using a J48 decision tree classifier, based on the likelihood of the Tweet being retweeted by a particular user. Although the authors address information relevance from a user-centric point of view, the validation of whether a prediction of a retweet occurring for a given Tweet is actually indicative of the *interestingness* of said Tweet do not perform particularly well.

A retweet-prediction model based on a factor graph model is introduced by Yang et al. [61] to determine how retweetable a Tweet is on a global scale. Although the methods are validated for predicting whether a particular Tweet will be retweeted, no mention is made of how this relates to how *interesting* the information is. Another study into retweet prediction was carried out by Zaman et al. [64], in which a trained probabilistic collaborative filter model (named 'Matchbox') was used to determine the useful features in making the predictions. As with the previous study, the research focuses on a retweet *probability*, which is a binary decision made by one particular user. The methodology is not aimed at the inference of interestingness, and simply determines that the most relevant features for accurate decision predictions are the author of the original Tweet and the retweeter.

Conversely, Suh et al. [52] and Hong et al. [28] predict the *type* of messages that are likely to be retweeted further, the latter using a logistic regression to both predict an individual retweet decision and a retweet *volume*. The methods do not apply these notions to how interesting the information actually is to a particular user, achieve low recall and the multi-classifications seems only to perform well on very unpopular or very popular Tweets. It is made clear, however, that the retweet volume of a Tweet is useful in denoting Tweet *popularity*.

An approach is made by Celebi and Uskudarli [14] for the purposes of recommending users to follow based on a user search. The methods rely on analyses of the content of users' Tweets and the similarity of these to those posted by other users, and produce an ordered list of recommended authors. Queries are enriched to form keywords, which are shown to perform well in retrieving more relevant users when the methods are validated in a user study. Although the processes and analyses are not significantly relevant to the work in this thesis, the content properties addressed by the authors are helpful in understanding Tweet and user features.

Petrovic et al. [47] use a passive-aggressive machine-learning algorithm to make binary predictions on retweet decisions and cited that social features - for example, number of followers of the author, frequency of Tweeting, etc. - were the largest factors in the performance, and Naveed et al. [43] use a logistic regression, partly using a dataset published as part of another paper by the same authors as Petrovic et al. [47], to predict retweet decisions in order to address information interestingness. However, little effort is made to define interestingness or, indeed, validate that the inferences towards this are accurate and correct. A logistic regression is again used by Zhu et al. [66] for predicting binary retweet behaviours with the focus on information propagation in disaster scenarios, and Peng et al. [46] showed that conditional random fields can perform better than logistic regressions when modelling retweet behaviour in the same way.

Since the above papers only effectively consider a prediction of retweet outcome, which is a binary decision, it is hard to relate this more to the notion of global interestingness, aside from stating that a retweet implies the retweeter's relative interest in the Tweet. However, a retweet count, as mentioned above, is inappropriate as an indicator of *magnitude* of interest, and so the research into predicting individual retweet decisions cannot be used as a basis for this. Additionally, not much emphasis is placed on how well the techniques work on a more 'on demand' basis; many of the methodologies discussed require several features that may take a long time to collect and compute, making them unsuitable for use as part of quick and useful interestingness evaluations.

The idea of Tweet scoring (through more semantic analysis) and retweet *count* predictions is introduced by Gransee et al. [25], who used their methodologies to produce a system[8] enabling users to compile Tweets in ways that are predicted to achieve the most retweets. The predictions are based on averaging the score, derived through a linear regression, of different components of a user's Tweets (such as the inclusion of a particular hashtag), so that when a Tweet by the same author is next constructed, the various components of the new Tweet can be compared against the scores of the counterparts seen in previous Tweets. The value produced through this method is then used to generate an expected retweet count as part of a comparison to the user's average ('baseline') achieved retweet count at this point in time, and was shown to perform well on influential Twitter users. However, the methods described do not take into account fluctuations in the social graph, particularly in the case of less-influential Twitter users, who's local networks are prone to more frequent changes. Additionally, they rely on a significant amount of previous Tweet and temporal information on the user to be evaluated, and do not relate the resultant score to any type of interestingness metric

[8] https://sites.google.com/site/learningtweetvalue/home

in the context of highlighting it from amongst noise.

Alonso et al. [4] also use ‘scoring’ to address interestingness, again focusing more on semantics through determining *uninteresting* content. In this work, Tweets are assigned an integer score out of five. Although the authors initially attempted to train a decision tree classifier on a set of 14 features, they begin classifying a Tweet as ‘possibly interesting’ if it contains a URL, and otherwise classify it as ‘not interesting’. Although the authors did then further classify the possibly interesting Tweets, by studying the magnitude of the crowdsourced participants used to evaluate the Tweets that found them interesting, and then classifying Tweets based on them containing a particular type of named entity (for example, a person’s name, a place or brand name, and so on) the categorisation system is too coarse and is not capable of representing the many different types of Tweets seen on Twitter. Additionally, despite achieving relatively high accuracy in this particular area, the methods are not suitable for assessing Tweets on a general or user-specific level, especially since Tweets that don’t contain URLs might still contain interesting content.

Alhadi et al. [1] and their later paper [2] introduce a system, called ‘LiveTweet’[9], which attempts to predict Tweet interestingness through applying scores based on retweet probability. The authors’ methods involve the continual updating of a model containing information on the features of Tweets that are being retweeted the most at a given point in time. Many of the features used are semantic, including term recognition and sentiment analysis, but other and more static features, such as inclusions of URLs, are also used. The methods are disadvantaged in that they rely on this continual re-building of the semantic model in order for the system to work. In addition, little indication is made by Alhadi et al. [2] of how accurate the probabilistic scoring is in terms of its agreement with users. The authors do agree, however, that a Tweet being retweeted cannot alone imply interestingness, due to factors such as user influence and time of day, but that a single retweet decision does imply that user’s interest in the Tweet.

[9]Available at: http://livetweet.west.uni-koblenz.de

An interesting study is described by Lauw et al. [36], in which a clustering algorithm is used, taking into account the retweet count of a Tweet and how this is related to the popularity of the source user, to determine information quality. Although this work is more similar to the research discussed later in this thesis than others, the scoring is quite simple and the author's use-case seems limited to that of identifying the most important Tweets surrounding a particular event (such as the death of Michael Jackson). Additionally, the authors do not make any effort to verify their results in any way, aside from comparing the Tweets determined to have a high quality by each of their two assessed methodologies.

2.2.6 Twitter is a 'Memepool'

A large amount of research in this field, particularly in the case of the work involving machine learning and classification, as discussed above, relies on feature selection and extraction. By choosing an appropriate set of features that are able to represent the entity more accurately, then this enables the model produced from the features to have a greater classification performance.

In 1976, Richard Dawkins coined the term 'meme' to be defined as a "unit of cultural transmission" [18]. The general idea behind memetics is as an analogy to biological genetics except, unlike genes, memes are entirely non-physical and represent a cultural idea or aspect or another human-based behaviour. The rise of social networks on the Internet has allowed the spread of memes to grow to the extent that they are sometimes now even represented by 'physical' constructs, such as images.

In genetics, a gene is a physical entity containing information and instructions. It is a unit of genetic inheritance, in that they are passed from parent to offspring through the act of reproduction, and the result of an organism having a gene is that the organism will express the features represented by that particular gene. These genes contain instructions that make up the features of an individual, such as physical characteristics like eye colour and height, and non-physical characteristics, including various aspects of personality.

Organisms exist in an environment that also has features, such as humidity, altitude, temperature, relationships to other organisms, and so on. If the genes of an organism are such that they cause the individual to be well-suited to its environment, then that organism has a better chance of survival and, therefore, a better chance of achieving reproduction.

Memes are similar in that they are effectively made up of a set of features, or a 'memome', such as the wordings of a particular phrase, or their relevance to other cultural aspects. These enable the meme to be less or more likely to be replicated in different environments, which is made up of the humans exposed to it and the interactions between them. For example, an Internet meme relating to the Star Wars movies would likely have a greater chance of being reproduced, through discussion and reposting, in an environment comprising a set of science-fiction fans than when amongst more mixed-interest groups.

The meme is also a useful analogy in this thesis when describing the way in which Tweets undergo replication within Twitter and for feature selection. Like a meme, a Tweet has a specific set of features, such as the text it contains, the inclusion of any mentions or a URL, and so on, and it exists within an environment consisting of a set of interconnected users on the Twitter social graph. A particular Tweet would generally have a greater chance of 'surviving' and being replicated, through the act of retweeting, amongst a certain subset of users interconnected in a particular way than in other environments.

As such, the Tweet features are analogous to the *genes* of a genome, and the arrangement and type of users on the social graph that receive the Tweet and have an opportunity to assist in its propagation comprise the Tweet's *environment*. Both of these aspects are of importance and are considered as part of feature selection in the relevant parts of this thesis.

2.2.7 Precision and Recall

Precision and recall are two metrics that are often used simultaneously to verify the performance of a method or procedure for information retrieval, with the usual goal being to maximise both. The metrics are used for validating *accuracy* in different ways, yet they can be applied to other purposes also and are useful in describing the notion of interestingness in Twitter.

"Classic" precision and recall are derived from the ratios of relevant documents to non-relevant documents and consider also the relevant documents that *aren't* retrieved by the system for a given search query. In particular, precision is the proportion of documents retrieved that are relevant, and recall is the proportion of relevant documents that were retrieved;

$$Precision = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of retrieved documents}}$$

$$Recall = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents available}}$$

The precision and recall measures have been useful tools in Twitter- and retweet-based literature. These pieces tend to only analyse the measures on their own work when applied to Twitter rather than on any more global scale. Certainly, there is less in the literature on the subjects of precision and recall with regards to retweeting in general.

The idea of assessing the credibility of information is introduced by Castillo et al. [13], who demonstrate methods of measuring the credibility of 'news' and 'chat' Tweets. In this case, retweeting is seen as a possible measure of a Tweet's credibility, since users typically only retweet information they see as interesting or useful. The authors use a logistic regression on a set of features derived from each Tweet in order to classify its credibility.

The precision and recall metrics are used to verify the different aspects of the paper's results. In particular, they are applied to the classification of assessing credible information (and users) in order to calculate how well classified the information is. A higher precision, therefore, shows that their model has accurately classified most of the total information classified as either credible or non-credible.

$$Precision = \frac{\text{Number of correct classifications}}{\text{Number of total classifications made}}$$

$$Recall = \frac{\text{Number of correct classifications}}{\text{Total number of potential classifications}}$$

On a similar note, Hong et al. [28] discuss the notions of precision and recall more generally. The authors discuss the problem regarding the balance of information received by Twitter users. Having too few friends reduces the number or interesting posts received (i.e. low recall); having too many friends may cause information overload and is likely to include a lot of noise (i.e. low precision). This issue is used, instead of for the purpose of validating results, as a basis for the work; predicting the Tweets that are most popular and will be retweeted the most.

In addition, precision and recall are used to compare the method to two other baselines; the TF-IDF score, which in this case is used to indicate how important the terms are in each Tweet; and *Retweet Before*, which uses the fact that if a Tweet in the training data has been previously retweeted, then it's likely to be retweeted again. The two metrics are also used to compare results when certain features are removed from the classifier. For example, showing that without using a 'user retweet' feature, the precision and recall remain significantly higher than when removing other features, meaning that this feature does not contribute highly to the performance. More specifically, precision and recall are used in a similar way by Castillo et al. [13]; except rather than looking at the number of classifications made, the authors use the number of predicted retweets.

[10] discusses a proof of concept for detecting influential users in one of two categories;

evangelists or detractors. Precision and recall, in this case, are used slightly differently:

$$Precision = \frac{\text{Number of influential users retrieved}}{\text{Number of users retrieved}}$$

$$Recall = \frac{\text{Number of influential users retrieved}}{\text{Total number of users}}$$

The concept is taken further through the use of another metric, the *Mean Average Precision*, which is used to denote an influential user as being a detractor or an evangelist. A high precision, in this case, would imply a large proportion of influential users are classified correctly and a high recall means that most of the influential users existing in the entire dataset have been classified. The final results then show the precision and recall values for detecting evangelists and detractors in both follower/following networks and interaction networks. Both precision and recall improved when the size of the set of highest classified influentials increased (i.e. the top set of influential users).

Pak and Paroubek [45] present a method for the automatic classification of Twitter information to determine if a document (or Tweet) is positive, negative or neutral in sentiment. In this case, the authors replace precision with *accuracy* and recall with *decision*, since they are using many classes instead of a binary classification, but use a very similar definition of accuracy (precision) to that used by Castillo et al. [13]. The decision is defined as:

$$Decision = \frac{\text{Number of retrieved documents}}{\text{Number of all documents}}$$

The accuracy is measured across the classifier's decision, and the $F_{0.5} - measure$ is then calculated based on these values instead in order to show that the classifier works well when the dataset size is increased.

As well as a news source, Twitter is also used as an informational, user-contributed source on world events. Marcus et al. [37] introduce a system, TwitInfo, which can be used for detecting, summarising and visualising events from Tweets. The authors looked at football match footage, web content, and earthquake survey data, and manually annotated major events in each to produce ground truth sets. These would be used to compare and contrast the results produced by their event detector using the following definitions of precision and recall:

$$Precision = \frac{\text{Number of events detected that are in ground truth set}}{\text{Total number of events}}$$

$$Recall = \frac{\text{Number of events detected}}{\text{Number of events in ground truth set}}$$

With these definitions set, the authors were then able to easily calculate precision and recall for their algorithm.

For the work in this thesis, interestingness of information is the performance metric used to describe information quality, and thus precision and recall for any particular user in the scope of this thesis can be defined as follows:

$$Precision = \frac{\text{Number of interesting Tweets received}}{\text{Total number of Tweets received}}$$

$$Recall = \frac{\text{Number of interesting Tweets received}}{\text{Total number of all interesting Tweets}}$$

where *received* means that the Tweet has arrived on the user's home timeline, but does not imply that the user has *read* the Tweet.

Therefore, a user following many other users will receive lots of interesting information onto their home timeline in amongst lots of noise; resulting in a reduced precision and higher recall. Another user might follow a very select few other users who are of direct interest, and thus will experience high precision, but low recall. These metrics are therefore useful in describing the concepts of noise and interestingness, and are consistent with their respective definitions in that users will achieve an optimum Twitter experience if both precision and recall are maximised.

Zadeh et al. [63] defined bespoke definitions of precision and recall, yet also in the domain of interesting information on Twitter. Although the authors identify the need for users to be able to discover other users of interest and declare that Twitter does, in fact, have a 'high precision' of interesting information, they admit to using a very coarse set of possible interest categories and is only based on *overlapping* interests rather than addressing the interest-noise ratio more concerning the research in this thesis. Additionally, clicks on URLs by users are the only means by which to measure this interestingness, and Tweets with URLs are usually the most interesting type of information [4].

2.3 Collecting Twitter Data

Most of the analytical work in this thesis relies on various data being collected from Twitter. Twitter provides an API for developers in order to facilitate the production of applications for its platform, but also for research purposes. It permits interfacing with many components of Twitter's service, such as posting and retrieving Tweets, interacting with other users (e.g. creating new friendships), and most of the features that Twitter's service itself provides to its users. The API encourages use of the OAuth[10] authorisation framework to handle access[11], allowing Twitter to keep track of applications and each application's access privileges and rate limits[12].

Twitter's traditional REST API, v1[13], provided many useful endpoints for data collection and allowed each OAuth-authenticated application 350 hourly POST and GET requests[14]. In June 2013 Twitter officially deprecated v1 of its REST API, forcing use

[10]http://oauth.net
[11]https://dev.twitter.com/docs/auth
[12]https://dev.twitter.com/docs/rate-limiting/1.1
[13]https://dev.twitter.com/docs/api/1
[14]https://dev.twitter.com/docs/rate-limiting/1

of its new v1.1 API[15]. The new version contains many of the same resources[16] as the original, but workarounds are required to get the same results as some of the endpoint requests possible through v1. Additionally, new rate-limit policies were introduced, allowing more limited and controlled access to most of the available resources.

Since the work in this thesis was ongoing over this switch-over date, the initial work utilised API v1, and the latter work API v1.1, causing some changes to some of the data-collection methodologies as the thesis progresses. Descriptions of the data-collection in each relevant part of the thesis reflect this change, where appropriate.

2.4 Motivation

The motivation for the research questions declared in the previous chapter lies in the need to distinguish interesting information from noisy Tweets in Twitter, the latter of which is the problem area identified over the previous sections of this thesis. Much of the most relevant research has studied retweet-prediction as a whole, without making ties between this and interestingness. Some of the authors have taken this further in order to develop methods for identifying interesting users and Tweets. A lot of this has involved the semantic analysis of Tweet content or term similarity, uses methods or models that must be actively maintained over time, or generated results that have not indicated strong performance or been rigorously validated in user studies.

This has provided motivation for research into a non-semantic method for quickly and accurately generating global interestingness inferences for given Tweets from a wide spectrum. Of importance is a model that can represent a large number of different users so that certain factors affecting retweetability, such as influence, can be abstracted away to allow all Tweets to be evaluated on a similar scale and to negate the need to be updated to suit the continuously changing user graph.

[15]https://dev.twitter.com/blog/api-v1-retirement-date-extended-to-june-11
[16]https://dev.twitter.com/docs/api/1.1

Throughout the review of the literature, it has also been made clear that the retweet count of a Tweet cannot reliably be used alone as a measure of interestingness, especially in the context of influential users. These users naturally achieve significantly more retweets than less-influential ones, but this does not imply that the information they produce is of a higher quality or interest level. As a result, the retweet count alone cannot be useful in distinguishing interesting information from noise in a timeline of mixed Tweets from different users with different levels of influence - some further metric is required to make this distinction. Thus, questions **RQ1** and **RQ2**, together forming part of the hypothesis in Section 1.3, have been partially addressed and have helped show why a method for estimating interestingness is useful. The remaining questions pose curious research pathways, which are introduced and explored further over the following chapters.

This thesis covers the research and development into a methodology for determining and ranking information on Twitter by inferred interestingness through considerations of user influence and the structure of users on the social graph.

In the following chapter, an understanding is provided on the structure of Tweets and of the underlying social graph of Twitter. Key definitions are explained to form the foundation of the research in later chapters. Ties between the audience of a Tweet and the propagation pathway made by retweets are drawn, which is important for understanding the social graph analyses explored in Chapter 4 and the different propagation characteristics observed across different network structures.

Chapter 3

Understanding The Behaviour of Retweeting in Twitter

It is clear how the popularity of a Tweet can be related to its propagation characteristics as it travels through Twitter's social structure. That is to say, that the greater the number of times a Tweet is retweeted by users, the more people have found its contents to be interesting enough to be worth sharing.

It has also been shown that this retweet count metric alone cannot be a direct implication of the actual interestingness level of a Tweet. Reasoning for this is related to the notion of user influence, which dictates that some Tweets are naturally immediately available to more people and thus have a higher chance of achieving a retweet. Indeed, Suh et al. [52] demonstrated that a user's Tweets' retweet rates increase as the user's follower count increases.

The strength of Twitter lies in its social structure, where users can elect to follow and unfollow others as they desire and with immediate effect. Followers of a user receive all of that user's posts onto their individual (or 'home') timelines. As a result, people are likely to follow users who create more 'interesting' posts; whether the follower is very interested in the friend his/herself, or if the follower is simply interested in the topical area of most of the friend's posts.

Just as Twitter users will post Tweets with subjects that are of interest to them - possibly related to the user's work, a hobby, or a mixture of multiple areas - and these Tweets are generally posted with the idea that they will be useful or interesting for some of their followers as well as an attempt to attract more followers, retweets are generated with the same motives in mind. This means that if a Tweet is retweeted, it is not only allowed to disseminate further through the social structure, but also that a higher Tweet quality is implied.

Thus, this describes how a user's friends, who carry out most of the retweets of the user, effectively become filters of interesting information for that user and for the followers of those friends. The *audience* of the original Tweet is therefore significantly increased. Since retweets are usually always attributed to the original author then you, a Twitter user, may gain more attention by means of followers by posting *interesting* Tweets, which will;

1. increase the chances that users reading your Tweets will choose to follow you, and;

2. increase the chances that users will decide to retweet your Tweet, thus broadcasting it to a larger audience. People viewing this *retweet* may then decide to follow you.

Since a Tweet can be retweeted multiple times, and that a retweet itself can also be retweeted, the much larger the effective audience (both directly and through retweets) of a Tweet's original author has the potential to become if they choose to post interesting information. In this chapter, an understanding of the behaviours and properties of retweets is provided, along with discussions into how these are relevant in researching useful metrics for determining which retweeted information is interesting.

The notion of 'global interest' is used in this thesis as a definition of interestingness. Although it is not feasible to say that a particular Tweet is unanimously interesting, it is possible to identify Tweets that are *generally* more outstanding than noise.

In this chapter, several contributions are made. Key terms, such as retweet group and path-length, are defined in the context of Twitter and Tweet propagation and become a model for much of the remaining work in the thesis; an in-depth analysis into retweet behaviour, including the dynamics of Tweet spread, audience, and temporal characteristics, is made; and discussions are conducted over the penetration of retweets and their relationships to communities on the social graph. Questions **RQ1** and **RQ2** from Section 1.3 are addressed, reinforcing the need for the non-semantic inference of globally interesting information.

Understanding these concepts is useful fow highlighting the problem area further, as they reinforce the disconnect between the graph produced by retweet propagation between users and the underlying social structure of Twitter. The interestingness model developed in Chapter 4 exploits this disconnect and the users' abilities to identify information that is interesting. The notation discussed in this chapter also helps support the mathematical definition of interestingness discussed later in Chapter 5.

3.1 Tweet and Retweet Properties

In this section, a more formal overview on retweet properties is provided in addition to an introduction and definition of concepts frequently referred to in the thesis.

3.1.1 Retweet Groups

A Tweet has various attributes associated with it, which make up the features that describe it and its author. These properties relate to the content, the author, and other metadata, such as its creation time, geographical location coordinates, language, and so on. However, not all of these properties are relevant to this research and, as such, a particular Tweet, t, has its relevant properties declared and defined as follows;

$$t = (\mathrm{text}, \mathrm{count}_R, \mathrm{author}_O, \mathrm{author}_R, \mathrm{orig}, \mathrm{prev})$$

Respectively, this represents the Tweet's text, its retweet count, and the *original* author of the Tweet. The final three values depend on whether t is a retweet or not and represent the author/forwarder of the retweet, a reference to the original Tweet, and a reference to the *previous* Tweet in the chain respectively, and are all `null` when t is not a retweet. Since a retweet is simply an extension of a class of Tweet, then the same properties can be assigned to retweets as to Tweets, except that in the case of retweets the values prev, orig and author_R will be non-`null`.

For example, let the Tweet shown in Figure 2.5 be t_1. This is an original Tweet and has the following properties;

$$\begin{aligned}
t_1.\text{prev} &= \texttt{null} \\
t_1.\text{orig} &= \texttt{null} \\
t_1.\text{author}_R &= \texttt{null} \\
t_1.\text{author}_O &= \texttt{Adrian Bradley} \\
t_1.\text{count}_R &= 0 \quad \text{(at time of writing)}
\end{aligned}$$

Inversely, the Tweet displayed in Figure 2.4 (r_1) *is* a retweet of another Tweet, t_2, where;

$$\begin{aligned}
r_1.\text{prev} &= t_2 \quad \text{(assumed)} \\
r_1.\text{orig} &= t_2 \\
r_1.\text{author}_R &= \texttt{Discover The World} \\
r_1.\text{author}_O &= \texttt{Alda Sigmundsdóttir} \\
t_2.\text{count}_R &= 48 \quad \text{(at time of writing)}
\end{aligned}$$

and $r_1.\text{count}_R$ is unknown. $r_1.\mathrm{prev}$ is assumed since r_1 was created using the button retweet method, which does not cite any intermediate retweeters, even if they exist.

A Twitter user, u, is represented by a Twitter account, and also has a set of properties. In relevance to the work in this thesis, these largely relate to the user's position in the social graph.

The full social graph, denoted by G, comprises $V(G)$, the nodes representing the set of users on Twitter; and $E(G)$, which is the set of edges connecting these nodes. In Twitter's case, the edges denote the followships between users, and are therefore *directional*. Thus, the set of followers and the set of friends of user $u \in V(G)$ are denoted by $N^+(u)$ and $N^-(u)$ respectively, where;

$$N^+(u) = \{u \in V(G) : \overrightarrow{uv} \in E(G)\}$$

$$N^-(u) = \{u \in V(G) : \overleftarrow{uv} \in E(G)\}$$

In other friendship-based social networks, such as Facebook, relationships are mutual and are therefore represented by non-directional edges in the relevant social graphs.

The terms $\text{deg}^+(u)$ and $\text{deg}^-(u)$ respectively refer to the in-degree and out-degree of a user u, where $u \in V(G)$. These in turn represent the cardinality of each of the sets of followers and friends of u, and therefore the author of Tweet t has a follower count of $\text{deg}^+(t.\text{author}_O)$ and a friend count of $\text{deg}^-(t.\text{author}_O)$.

Let T represent the set of *all* Tweets. Since a Tweet can be retweeted more than once, and have its retweets also retweeted, the set of retweets of Tweet $t \in T$ is defined as;

$$RT(t) = \{r \in T : r.\text{orig} = t\}$$

Hence, the retweet count of t is given by $t.\text{count}_R = |RT(t)|$.

Definition 3.1

A **retweet group**, denoted by $RG(t)$, describes the original Tweet, t, along with the set of the retweets of t, $RT(t)$. Thus;

$$RG(t) = \{t\} \cup RT(t)$$

Retweet groups are useful for identifying a Tweet and the retweet replicas of it, and is appropriate when discussing the audience reach of a particular Tweet. Therefore, since t is also a member of this set, the size of t's retweet group is;

$$|RG(t)| = t.\text{count}_R + 1$$

which can have a minimum cardinality of one - $RG(t) = \{t\}$ - in cases when not retweeted at all.

3.1.2 Retweet Trees

As a Tweet gains popularity and is retweeted more, and since its retweets themselves can *also* be retweeted, then this results in the generation of a retweet *tree*, which represents the retweet group of a particular Tweet. This tree illustrates the original Tweet and the various propagation pathways it takes as it is retweeted through the social graph.

The tree is not a representation of the actual social ties between the authors of the tree's nodes, as users are able to retweet Tweets and retweets sent from others that they do not follow. However, as is mentioned later in this chapter, most retweeting does generally occur between directly-linked users. [35] also uses retweet trees to assist in illustrating information dissemination in Twitter, particularly in observing the Twitter reactions to the 2009 Air France airline crash.

The root of the tree is t and, if t has been retweeted, each of the other nodes are made up of the set of retweets in $RT(t)$. Each non-root member of the tree refers to its parent through its own 'prev' attribute, as illustrated in Figure 3.1. Retweet trees are useful for this purpose as they help demonstrate the temporal 'paths' down which the retweets occur and the chains they produce. A similar illustrative device is used by Galuba et al. [22] in describing URL cascades in Twitter.

In very rare cases, more than one node in a retweet tree may share an author user. This only occurs when a particular user retweets a Tweet more than once, and would only generally happen in scenarios where the user is using the manual method to modify the content as part of a conversation with others or for expressing multiple opinions. For example, a user may receive a Tweet relating to a particular news story, and then decide to retweet it with a small annotation. Upon feedback from followers, the user then retweets the Tweet again, yet with a different annotation. Each of these new

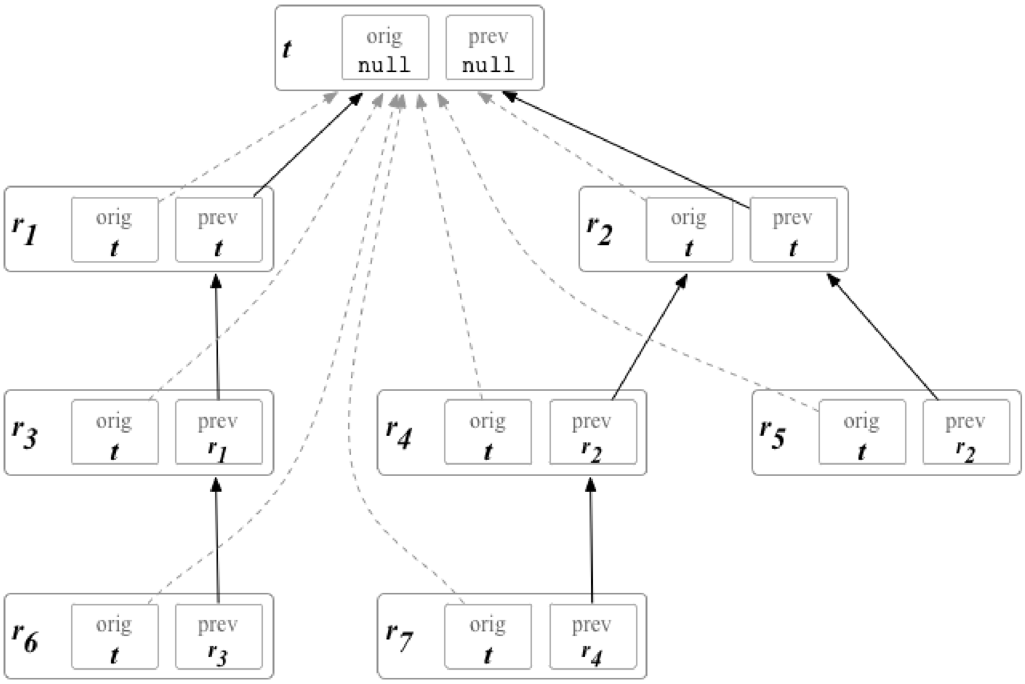


Figure 3.1: A hypothetical retweet tree.

retweets could then become the root of two branches in the complete tree of the Tweet.

Retweeting the same Tweet multiple times is not supported through the button method. Once a Tweet has been retweeted by a user in this way, there is no provision for the functionality to retweet a different member of the same group or, indeed, the original Tweet. A retweet can be 'un-done' by clicking the button again on any member of the retweet group, yet this will not affect further retweets of this retweet that have been made using the button method.

3.1.3 Path-Length

In addition to retweet groups having a size property, a retweet groups's branch's *path-length* refers to the length of a particular retweet chain.

Definition 3.2

The **path-length** of a single retweet chain in a retweet group is defined as the number of hops between a Tweet, t, and the retweet represented by the leaf node

of the chain's branch in $RG(t)$'s tree.
The **maximum path-length** of a retweet group is the greatest path-length observed in the retweet group.

Figure 3.1 represents the members of the retweet group of a hypothetical Tweet, t. This retweet group has a size of 8 and has 3 distinct retweet chains, the longest of which are the two involving $[t, r_1, r_3, r_6]$ and $[t, r_2, r_4, r_7]$. The *maximum* path-length of this retweet group is therefore 3, as the leaf node of both of these branches is three hops away from the original Tweet at the root.

Although the tree does not illustrate the edges between users on the social graph, it is possible for the underlying graph to connect the authors of the tree's nodes in various ways. For example, it is likely that $r_3.\text{author}_R$ follows $r_1.\text{author}_R$, but it's also possible that $r_3.\text{author}_R$ follows $r_2.\text{author}_R$. More on this topic is discussed in the audience analysis in Section 3.3.4 and in the social graph analyses in Section 3.3.5.

When a user retweets a Tweet or retweet through the manual approach, the current test of the Tweet is pre-pended with the sequence `RT @<username>:`. Therefore, a Tweet with the content;

`RT @user2: RT @user1: This is the body of the Tweet`

was originally authored by `user1`, then retweeted by `user2`, and then finally retweeted by the author of this current retweet (a Tweet or retweet's author's username is not credited in the body of the text in this way). Making such citations does count towards the 140 character Tweet limit, which may partially explain the path-length distribution pattern observed later in Section 3.3.2.

It should be noted that this phenomenon can only be observed through retweets by the manual approach, since the button method always simply credits the original author, and not any of the internal members of the retweet group. Although a significant number of retweets today are carried out using the button method, the manual approach still remains popular currently and even more so at the time the research in this chapter was carried out in the spring of 2011. This allowed for making useful observations of retweet patterns that could not be as successful later on.

3.2 Twitter Propagation Analysis

Understanding information propagation in Twitter is key to also understanding how interesting information might be detected. Whilst it is known that the retweet count of a Tweet cannot be used alone in inferring interestingness, since this is simply a level of popularity tied in with the author user's influence, it is still a factor in that users are more likely to retweet interesting information than noise.

Of particular interest is to achieve an overview of propagation behaviours in Twitter; the patterns in the properties of retweet groups, such as their sizes and penetration depth, temporal aspects of retweets and information on the social structure of Twitter itself with regards to propagation within it.

The remainder of this chapter involves an exploratory study of the retweet characteristics in Twitter to provide a further background, and which demonstrates the area's relevance towards the goal of inferring interesting information.

3.3 Retweet and Retweet Group Analysis

To assist in providing a further grounding in this area of research, a series of analyses were carried out into retweets and retweet groups. This section describes the processes and purposes of the analyses.

3.3.1 Data Collection Methodology

The analyses involve the examination of Tweets extracted from Twitter's REST API v1, which was used between 26th January and 24th May 2011 to collect Tweets and retweets from the public timeline.

The data collection involved a mixture of using Twitter's timelines and its search capabilities. Version 1 of the REST API supported retrieval of Tweets, 20 at a time, from the Twitter *public* timeline. Historically, this timeline contained the 20 most recent Tweets published by all the authors that have non-protected Twitter accounts, and it used to be visible on their website's homepage[1] to non-logged-in users.

In particular, for the data-collection periods, the public timeline endpoint was queried every ten seconds to retrieve the current set of the most recent public Tweets. Millions of Tweets are posted each hour, and ten seconds was a granular-enough frequency to ensure that there was no duplication in the data returned. From all of the retrieved Tweets, the ones that were retweets were filtered out and stored. Retweets, as mentioned earlier, are distinguishable since they start with the characters 'RT' followed by a username. It should be noted that when retrieving Tweets from Twitter's API that even retweets that were created using the button method begin with the same character sequence, allowing detection of these also.

Following storage, the content of the retweets were parsed in order to extract the text that the original Tweet contained. Sometimes, retweets using the manual approach are used to provide additional annotation to the Tweet. Although this can often be distinguished by the fact that the original Tweet content is inside quotation marks (" "), it is not true in all cases, meaning that sometimes the original text could not be reliably extracted programmatically by a machine. In these cases additional queries were made to Twitter's search API in an attempt to resolve the problem, yet, failing that, the retweet was discarded.

[1]http://twitter.com

Once the original text had been successfully extracted, this was used along with other metadata as query parameters to Twitter's search API in order to try and find the original Tweet and any other retweets of this Tweet. The search API uses approximate (or 'fuzzy') string matching, but quotation marks can be used to retrieve search results based on an exact string pattern[2].

Once the API search was complete (in some cases, with Tweets achieving many retweets, many API calls were required in order to page through results), the original Tweet could easily be identified as the only one of the set *not* starting with the sequence "RT". This provided a retweet group comprising the original Tweet and all available retweets of this Tweet.

On some occasions, more than one Tweet were each identified as the original Tweet and so no data was stored for this group. This could occur, for example, if many users Tweet exactly the same text coming from an external source, such as a news webpage, and means that the entire set of retrieved Tweets are not likely to be part of the same retweet group. In cases where no results were returned, the retweet was not stored and was assumed to be an orphan retweet, perhaps as a result of a retweet of a Tweet posted by a protected Twitter account. Where no original Tweet could be identified it was sometimes possible to identify it through cross-matching against other retweets in the retrieved retweet group, but were discarded if unsuccessful.

The retweet groups were finally stored along with relevant metadata in order to carry out the studies described in the following sections. The data consisted of a set of around 4,400 retweet groups, which comprised of 26,000 Tweets (defined as T') and retweets. The relatively limited size of the dataset is acknowledged, yet it should be emphasised that these analyses are simply exploratory and are not used to answer or solve any specific problem.

[2]https://dev.twitter.com/docs/using-search

3.3.2 Exploring Retweet Group Path-Lengths

The path-lengths of each chain in a retweet group can be calculated by identifying the members involved in retweet activity down that chain; from the original Tweet to the final retweet. The *maximum* path-length of a particular retweet group is the longest path-length observed in the group's tree.

Identification of path-lengths can be carried out through parsing the text of a retweet, and following the citations. Although it cannot be guaranteed that all users will be properly cited in a chain, and there is no realistic method to verify this, it is felt that correct citations will be made enough times to make these cases relatively insignificant.

On average, the maximum path-length observed across the retweet groups was around 1.8, with the vast majority of retweet chains being between one and two edges in length. When one considers that many retweets are made through the button method, which removes citations of internal users in the chain and simply credits the original author and would therefore produce many single-length retweet chains, this average will theoretically be an underestimate. The similar observations made by Kwak et al. [35] in the area also indicate a large number of groups with maximum path-lengths of one and two.

The longest observed maximum path-length was nine, which is a huge depth of penetration through the social structure since the total number of users involved in propagating the Tweet was ten. This, combined with the knowledge that social networks can represent a 'closer' social graph than the real world's six degrees of separation, shows how retweeting can have a huge impact in information spread amongst millions of people worldwide very quickly. Figure 3.2 illustrates the distribution of maximum path-lengths observed in the retweet groups of Tweets in T'.

Observation 3.1

The mean maximum path-length observed across the retweet groups analysed was around 1.8.

The longest maximum path-length observed was 9.

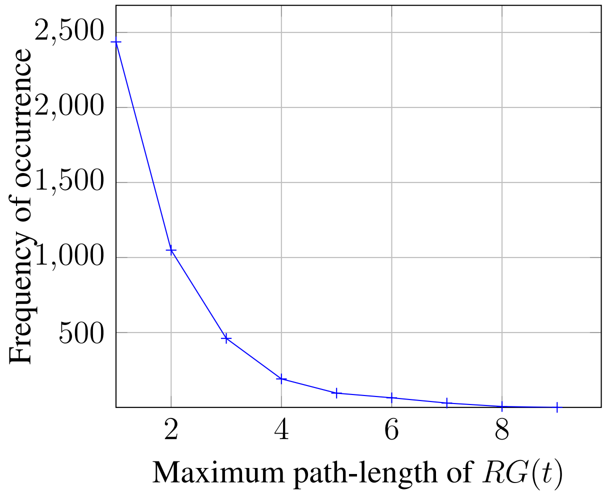


Figure 3.2: Distribution of maximum path-lengths observed in $RG(t)\,\forall\, t \in T'$**.**

Also of interest is the relationship in terms of the social ties between the different authors of the Tweets in a retweet group. In cases where a retweet group's maximum path-length is precisely one, i.e. the situation where a user (or set of) has retweeted a particular Tweet only once, the retweeting authors of the leaf Tweets of this group's retweet tree follow the original author around 90% of the time.

This implies, therefore, that in the remaining 10% of cases, a retweeter has retweeted a Tweet from outside of their home timeline and has instead seen a Tweet whilst browsing through another user, who isn't a friend, timeline that the retweeter regards as sufficiently interesting. This helps to demonstrate that the more followers a particular user has, the greater the chance that another user somewhere has of viewing the user's Tweets and then having the opportunity to retweet them. The fact that 90% of retweets of a particular user are created by direct followers reinforces this further.

This particular property could also be partly due to use of the button method of retweeting, which does not cite intermediate retweeters, and thus always imply that the final retweeter directly retweeted the Tweet from the original author. However, there may, in fact, have been other retweeters in between the final retweeters and original author, each of which following the immediately upstream retweeter. As such, this 90% follow probability between the retweeter and source user in 1-hop retweet chains is also likely to be an underestimate.

Further to this, in situations in which the maximum path-length of a retweet group is *greater* than one, retweeting authors in the group follow the author of the original Tweet about 40% of the time. It is clear from Figure 3.4 that retweet groups with a longer maximum path-length tend to have a larger size themselves. This increases the likelihood that the Tweet has been able to spread both further around the original Tweet's author's community, and also the potential for the Tweet to 'travel' to other communities. Since users from outside the source user's community are less likely to follow the source user, this explains the reduction in the followship likelihood between further downstream retweeters in the retweet chains and the original author.

3.3.3 Size of Retweet Groups

The distribution of retweet group sizes $\forall t \in T'$ was found to follow a power-law type distribution, with a relatively large p-value of around 0.87. Figure 3.3 represents the complementary distribution function demonstrating the changing probability of a randomly generated X being greater than or equal to x, the 'current' value of $|RG(t)|$, at each stage. The techniques used in this analysis are adapted from the methods and code provided by Clauset et al. [17].

The mean group size from this dataset was found to be just below six, and the largest size was 284. The smallest $|RG(t)|$ were the cases in which $t.\text{count}_R = 1$, and which were significantly the most common occurrences.

Observation 3.2

The mean observed retweet group in the dataset T' was of size 6.

Of interest also is the relationship between a group's size and its maximum path-length. Generally, the maximum path-length of a group, $RG(t)$, increases with $|RG(t)|$, indic-

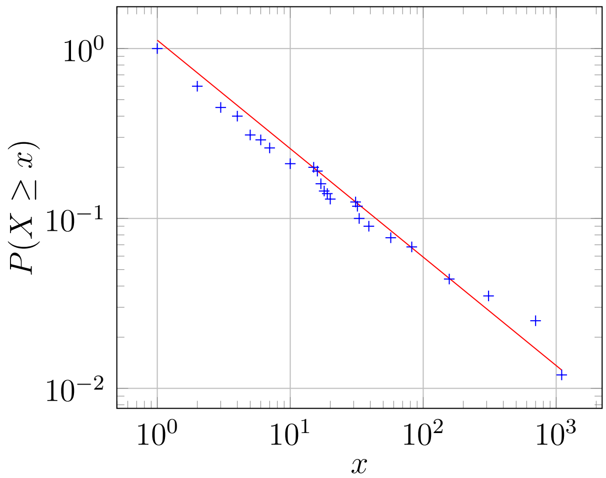


Figure 3.3: Maximum likelihood power-law fit for the cumulative distribution of retweet group sizes.

ating a mostly uniform growth in the retweet trees representing these groups - as might be expected. Figure 3.4 demonstrates this trend, which illustrates that as the retweet count of t increases, then the longer the retweet chains in $RG(t)$ are likely to become. This would increase its penetrative dissemination away from the source and further facilitate its spread between communities, increasing its potential *audience size*.

3.3.4 A Tweet's Audience - How Many Users Can be Reached?

$RG(t)$'s (immediate) audience size refers to the number of Twitter users that have received t, either in its original form or as a retweet, r, such that $r.\text{orig} = t$. Users in the audience are not guaranteed to have read the Tweets, but they are the users who will have received the Tweet on their home timelines.

The term 'immediate' is used to signify the distinction between those users who passively receive the Tweet, due to following the original author or a retweeter, and those who see the Tweet whilst actively browsing through other user timelines or the public timeline. Users in the latter group are therefore not direct followers of $t.\text{author}_O$ or $r.\text{author}_R \, \forall \, r \in RT(t)$ and thus cannot be tracked as members of t's audience.

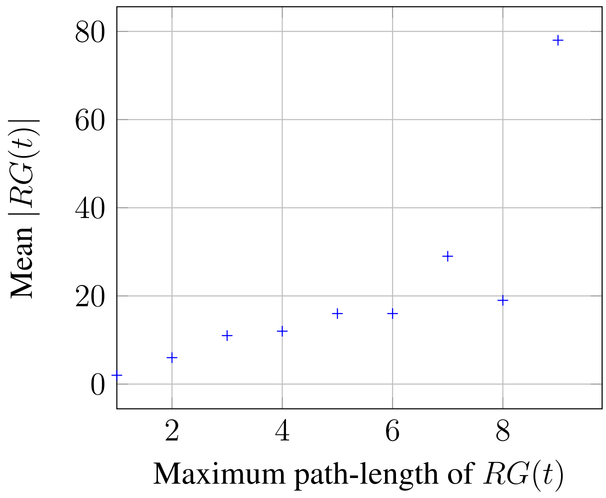


Figure 3.4: Relationship between the maximum path-length and size of a retweet group. The greatest path-length was included for context, but had a sample size of only one.

Let $r_1, ..., r_n$ be the members of $RT(t)$. The size of the audience of $RG(t)$ can then be calculated thus (assuming $t.\text{count}_R \geq 1$);

$$|\text{audience}(RG(t))| = |N^+(t.\text{author}_O)| + |N^+(t.\text{author}_r^1)| + ... + |N^+(t.\text{author}_r^n)|$$

where $t.\text{author}_r^1 ... t.\text{author}_r^n$ are the users who retweeted t.

Despite this, properties of Twitter's social graph dictates that this audience size calculation is naïve in that, particularly in the case of more tightly-knit communities, users who are authors of t or $r \in RT(t)$ are likely to share a subset of each of their followers. The more dense the communities, the more followers are likely to be shared between the authors in $RG(t)$ and, as such, the aforementioned audience calculation is likely to be an overestimate in nearly all cases. As such, audience$(RG(t))$ is a list and not a formal *set* of users, since it is likely to have some non-distinct members.

The following analyses of retweet group audience sizes relies on a dataset which began collecting at a later date than the general set used in this chapter, and thus the data represented in the rest of this section contains 2860 of the total 4400 groups originally collected. The longest maximum path-length of retweet groups observed in this subset was eight.

The *overhead* of a group, $RG(t)$, which attempts to address this problem, is related to the level of redundancy of received information by the audience.

Definition 3.3

$RG(t)$'s **overhead** is a value equal to the number of users in audience$(RG(t))$ that receive t or any $r \in RT(t)$ more than once and, if received more than once, the number of times t is received by each user.

The audience overhead was found to exist (be greater than 0) in 71% of all observed retweet groups, further reinforcing that retweets often occur within communities containing users sharing many edges.

Therefore, the actual audience of a Tweet is given by the *set* of users that can be found by modifying the existing definition to take the overhead into account;

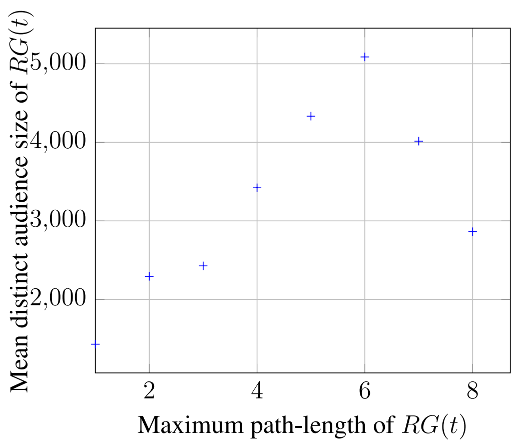
$$\text{audience}(RG(t)) = N^{+}(t.\text{author}_O) \cup N^{+}(t.\text{author}_r^1) \cup ... \cup N^{+}(t.\text{author}_r^n)$$

where $t.\text{author}_r^1, ..., t.\text{author}_r^n$ are the users who retweeted t. This definition ('distinct') of audience is used in preference over the previous ('raw') definition for the remainder of the thesis.

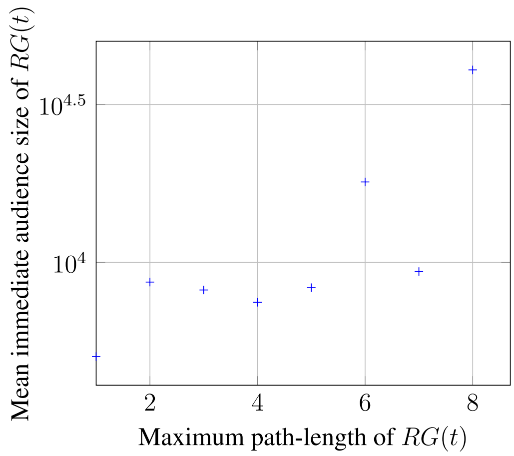
The *proportionate* overhead is defined as the ratio of the overhead to the audience size, and is sometimes more useful for analysing the size of the overhead compared to the popularity of the original Tweet.

For example, a Tweet t has been received onto the home timelines of 800 users as a result of a single retweet, r_1. 400 of those users received the Tweet twice due to the presence of shared followers between $t.\text{author}_O$ and $r_1.\text{author}_R$. In this case, the overhead is 400, the proportionate overhead is 0.5, the 'raw' audience has a size of 1200, and the 'distinct' audience has a size of 800.

Figure 3.5(a) illustrates, initially, that which might be expected; that the distinct audience size of a Tweet, t, is mostly proportional to the maximum path length of $RG(t)$.



(a) Varying $RG(t)$'s *distinct* audience size with its longest path-length.



(b) Varying $RG(t)$'s *raw* audience size with its longest path-length.

Figure 3.5: Comparison of the relationships between a $RG(t)$'s distinct and raw audience size and its maximum path-length $\forall\, t \in T'$.

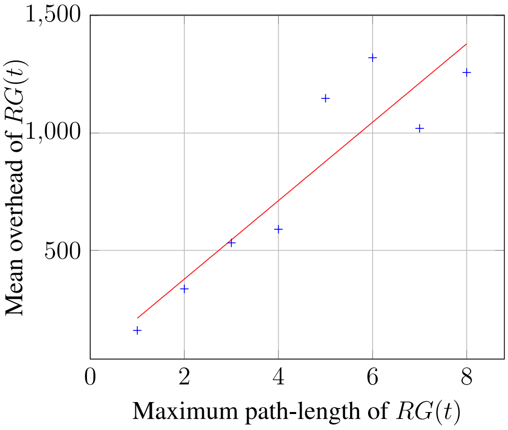
However, as the maximum path-length of retweet groups exceeds 5, then a *decline* in the distinct audience size is observed. This particular behaviour has an unclear cause, but it is felt that this could be related to the saturation of the proportionate overhead's ratio at this stage - in particular, that retweet groups attracting many retweets are circulated more within communities than outside and between communities. At this stage, the overhead becomes so large, causing this reduction in audience size. This is significant in that the distribution of the non-distinct over the increasing path-lengths demonstrates, mostly, a continuous positive correlation (Figure 3.5(b)).

The *largest* overhead was of a size over six times greater than the group's distinct audience size itself, indicating a massive overlap between the followers of the author of the original Tweet and the authors of its retweets. Whilst the audience overhead was only found to be greater than the distinct audience size in around 3% of observed retweet groups, it is still clear that the potential for overlap in the followers of retweet group members can be very large in more closely-knit communities Figures 3.6(b) and 3.6(d) show that the largest overheads observed diminishes in groups with greater maximum path-lengths, which helps illustrate this concept.

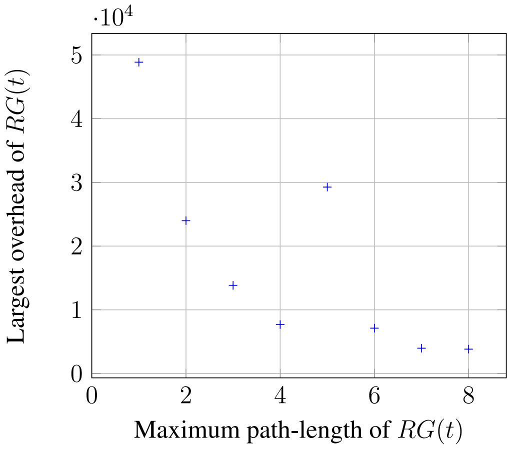
Conversely, Figures 3.6(a) and 3.6(c) respectively show that the *mean* overhead and mean proportionate overhead increase in retweet groups with greater maximum path-lengths. It is assumed that with larger retweet groups there is a greater chance for overlap between the followers of the authors of the members due to there being a greater audience size. Since it is known that increases in the sizes of groups can be indicated by increases in the groups' maximum path-lengths, then this suggests that, on average, the overhead should increase with maximum path-length.

The diminishing behaviour observed in the other two previously-discussed subplots suggest that these groups with smaller maximum path-lengths exhibiting greater overheads are those that do not fit the trends across retweet group size and maximum path-length observed earlier in this chapter. As such, it is likely that these groups are actually large, with representative trees that are shallow and very *wide*. This illustrates that the Tweets have not propagated far from the original author, yet have circulated thoroughly through a local community. Indeed, three of the largest five overheads in the set of analysed Tweets occur in retweet groups which have a maximum path-length of one.

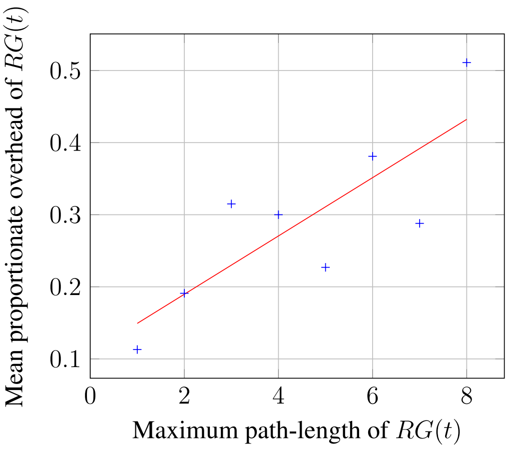
The power of the retweet phenomenon in terms of how it affects the potential audience reach of a particular Tweet is discussed in further detail by Kwak et al. [35], in which they find that a retweeted Tweet of sufficient interest can reach a very large number of users even if the original author has only a few followers. The same paper more specifically mentions that the audience size of a retweeted Tweet reaches, on average, at least 1,000 users, no matter the number of followers of the original author. This result agrees with the results in Figure 3.6(a) in that even Tweets with a short maximum path-length still, on average, have a relatively large audience size.



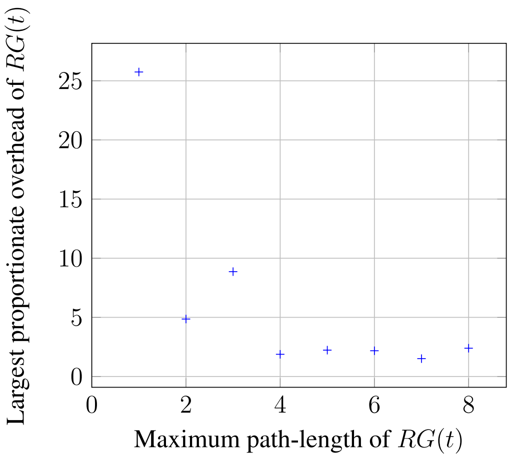
(a) Varying mean overhead with maximum path-length.



(b) Varying *largest* observed overhead with maximum path-length.



(c) Varying mean overhead *proportion* with maximum path-length.



(d) Varying *largest* observed overhead *proportion* with maximum path-length.

Figure 3.6: Relationships between $RG(t)$'s audience overhead properties and its maximum path-length $\forall\, t \in T'$, where T' is the set of analysed Tweets.

3.3.5 Retweet Groups on the Social Graph

Now that an understanding has been achieved in the behaviours and properties of retweets and retweet groups, it is important that the nature of the social ties between users in groups is studied. This will provide a grounding for the research in the following chapter, in which the social structure and its role in facilitating propagation, are discussed in more detail.

It has already been mentioned that the probability of a retweeting author following the original author in unit-length retweet chains was found to be around 90%. However, in retweet groups with longer chains, a decrease in the likelihood of the final retweeter (the user at the bottom of the retweet tree) following the original author was observed. Indeed, on average across all retweet groups, the final retweeter in the longest chain follows the *previous* retweeter in around 67% of cases. The final retweeter of a retweet chain is defined as the author of the leaf node of the chain.

It is interesting that this value should be about 20% lower than in unit-length maximum path-length groups, and it suggests that users have a greater chance of 'stumbling over' retweets found on non-friends' timelines whilst browsing through other users. Since it has been shown that with an increase in maximum path-length an increase in the audience size is also observed, then this demonstrates the increased chance of discovery of the Tweet through users searching through others' profiles. In cases where the maximum path-length of $RG(t)$ is equal to one, then the audience size is far smaller and thus there is a lower chance of users who aren't followers of $t.\text{author}_O$ or $\{r.\text{author}_R \forall r \in RT(t)\}$ finding the Tweet.

In addition, there is some evidence of user influence playing a role in the analyses of these data. In particular, in the 67% of retweet groups in which the final retweeter *does* follow the author of the retweet (or original Tweet) directly 'upstream', the latter user has, on average, around 950 followers. Inversely, in the remaining 33% of groups (in which the author of the final retweet does *not* follow the preceding author), the preceding author has an average of 600 followers. This is illustrated by an example in Figure 3.7 and it implies that there is a significant difference in the retweet potential with varying author influence levels.

This is further accentuated when one studies the follower connections of $t.\text{author}_O$ in groups where the maximum path-length is greater than one. Whilst it was found earlier that the likelihood of a $r.\text{author}_R$ following $t.\text{author}_O$ is around 40%, the average follower count of $t.\text{author}_O$ has a four-fold increase (from about 550 to 2,000) when

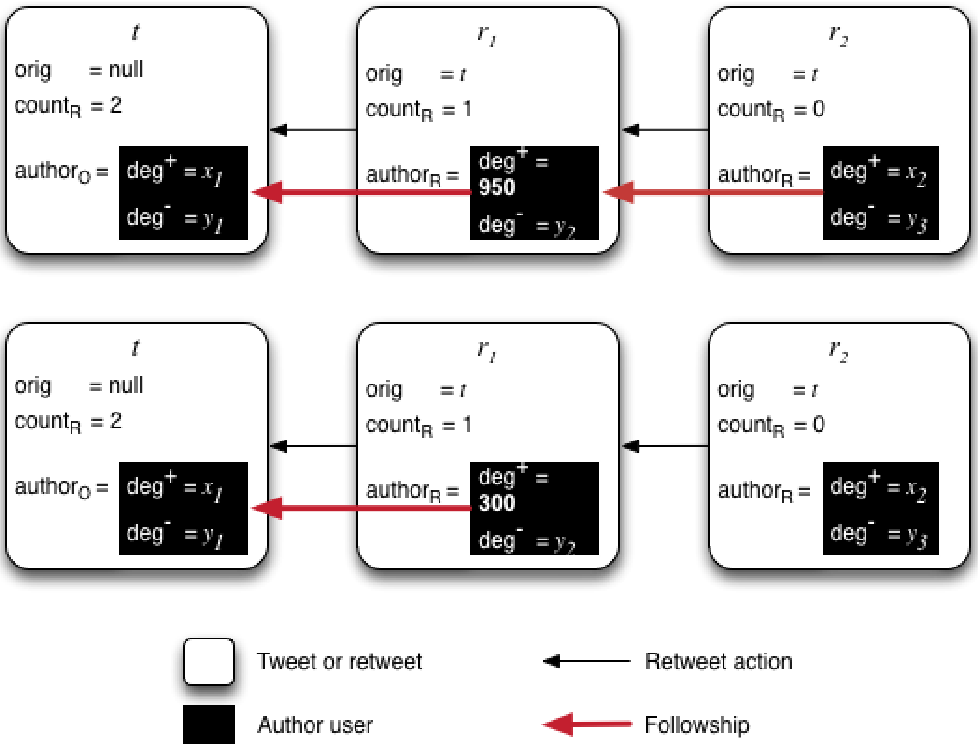


Figure 3.7: Effect of the final retweeter following the upstream user on the follower count of the upstream user.

he/she is also followed by the final retweeter. Figure 3.8 demonstrates an example of this in a retweet chain of two retweets; that when the author of r_3 follows the author of t, then $\text{deg}^+(t.\text{author}_O)$ is significantly greater.

In fact, in groups of *all* maximum path-lengths, $t.\text{author}_O$ had a consistently higher follower count when followed also by the final retweeter of $RG(t)$ than when not followed.

This particular behaviour also helps illustrate that a user is more likely to be retweeted when s/he has more followers - in this case, having four times the follower count increases the correlation dramatically (40% to 90%). The follower count can, therefore, be directly related in this way to the discussions of user influence by Cha et al. [15], and also of users using retweeted Tweets to passively ‘advertise’ themselves.

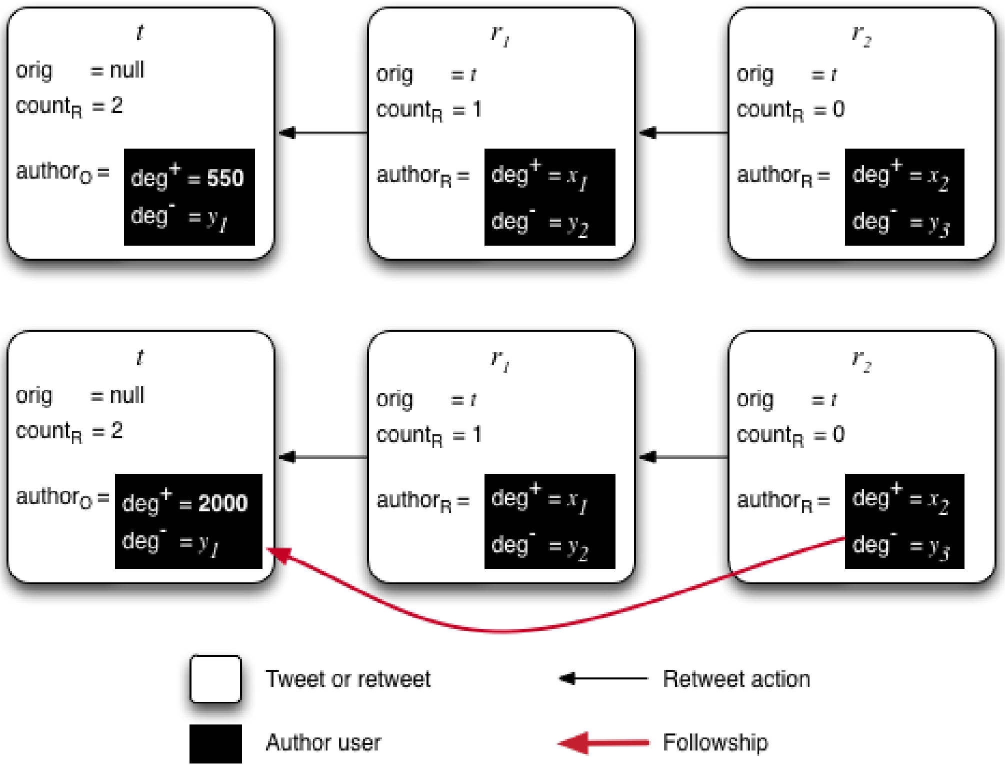


Figure 3.8: Effect of the final retweeter following the original author on the follower count of the original author.

Strangely, Figure 3.10 illustrates how increases in maximum path-length of retweet groups caused the follower count of the original author to diminish, indicating further penetrative depth of propagation when the original author has *fewer* followers. The collected retweet groups that contained longer retweet chains often also contained retweet chains that were much shorter. For example, a group containing chains with path-lengths of five, or more, are also likely to contain many more chains with path-lengths of one and two (as is implied in the distribution in Figure 3.2). There are, therefore, various possible explanations for this property, including the argument that users with many followers are generally likely to be part of a large community of users, from which retweets are not transmitted. Users that are part of several communities, and are therefore less involved with any given one, may find that their Tweets have the potential to be retweeted a further distance.

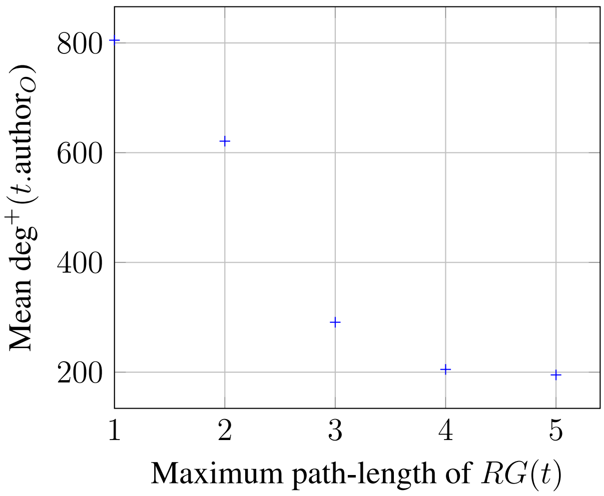


Figure 3.9: Analysis of variance in deg$^+(t.$author$_O)$ as $RG(t)$'s maximum path-length increases.

Additionally, and more interestingly, it is possible that users possess some awareness of their local networks and the users within them. A user, who is part of a large community with lots of obvious follower overlaps occurring between the members, may decide *not* to retweet a particular Tweet if he/she feels that many of his/her own followers may be shared with the author and that they might have therefore already seen the Tweet.

A final analysis on the social ties between users in retweet chains is carried out on the followship pattern of authors throughout the chain. Let h be the number of hops (or edges in the retweet tree) between the original author and a retweeter in a retweet chain. It was illustrated in earlier sections that, when $h = 1$, the likelihood of a retweeter following the original author is around 67%. However, as h is increased, then the followship likelihood mostly consistently decreases.

Let r_h.author$_R$ be the author of the retweet h hops from t in $RG(t)$'s retweet tree. Figure 3.10 illustrates how longer retweet chains do indeed increase both the likelihood of the Tweet reaching further through the social structure and the chance of achieving a smaller proportionate overhead.

Further to this, of the 67% of retweeters who *do* follow the original author when $h = 1$, only 19% follow also the upstream author at $h = 2$. In these cases, the latter has an observed average of around 3,000 followers. In the 81% of cases when the user at $h = 2$ *isn't* also followed, then the upstream author has a much lower average follower count of about 520.

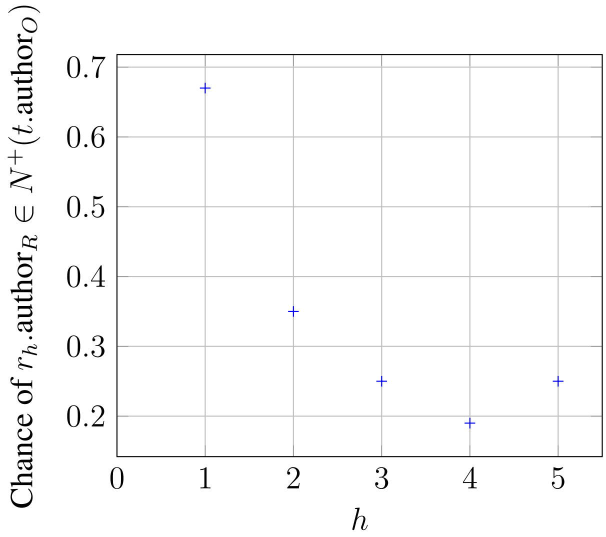


Figure 3.10: Relationship between the likelihood of r_h.author$_R$ $\in$ t.author$_O$ (where $r \in RT(t)$) and increases in 'distance' between r and t given by h.

It is, therefore, sensible to assume from these analyses that Tweets are forwarded more between groups of less-connected users, highlighting the notions of social network awareness and of community-hopping. If retweets were usually circulated around more closely-knit communities of users, then the followship likelihoods would be generally greater, more uniform, and consistent throughout the retweet chain. Users would have as much of a chance of following their immediate upstream neighbour author in the retweet chain as they would an author further upstream.

As mentioned near the start of this chapter, the author of the original Tweet should be cited by the `RT @<username>` sequence observed *closest* to the retweet body, where the `<username>` is the username of the original author user. Rather than specifically looking for the author's Tweet appearing in this location, Tweets were examined to check for the existence of the author's username being mentioned *anywhere* in the Tweet content, and was found to exist in about 68% of Tweets.

This frequency did not vary with any consistent correlation upon changes to the maximum path-length or retweet group size, and so it is assumed that users do feel the need to credit the original author more so than not.

3.3.6 The Temporal Properties of Retweets

The final set of analyses in this chapter relate to time's influence on retweet propagation. This provides insights into how quickly information can spread and, when combined with the knowledge of the social structure and audience, how this can relate to the rate of information dissemination and consumption.

Generally, it was found that the elapsed time between the original Tweet and the final retweet in retweet groups increased with the groups' maximum path-lengths, indicating that if there are more hops for a Tweet to travel down between users then it takes longer to do so. However, this correlation is only really applicable to shorter retweet chains, which more uniformly increase in elapsed time with increases in maximum path-length in a linear fashion roughly proportional to $v = \frac{s}{t}$, where the distance, s, is the hypothetical distance given by the number of hops between users, thus indicating that the speed, v, of propagation remains relatively constant.

Retweet groups exhibiting longer maximum path-lengths are less consistent in terms of the groups' propagation speeds. Whilst this is likely attributed to smaller samples, there are conflicting arguments for patterns observed in this propagation speed, which rely on various intervening factors.

As mentioned, the time taken for a Tweet to reach a specific path-length could be a function of the path-length itself, where as the path-length increases, then so does the time taken for the Tweet to be retweeted to the end of the chain. Inversely, Tweets that are especially popular, possibly as a result of being particularly topical (such as in the disaster cases mentioned in the Introduction), may be retweeted more quickly by users so that the information is spread more quickly. In these cases, retweet groups with longer retweet chains may complete their trees more quickly than those groups with much shallower retweet trees.

Similarly, user influence could play a role in dissemination speed; if a Tweet is retweeted by a user with many followers, then there is an increased likelihood of propagation through this user. Whilst this could, in addition to the previous argument, cause longer retweet trees to be completed more quickly than groups with shorter trees, it could also facilitate 'faster branches', in which particular long branches grow faster and reach their leaves more quickly than shorter ones in the same retweet tree, if the other branches consist of less-influential authors and retweeters.

There is not enough evidence provided in this analysis to make any inferences towards a generic pattern of retweet group growth speed, and it is believed that this growth is governed by many more factors than the Tweet itself or the social structure alone. As such, there is no predefined rule for predicting the spread of dissemination in this way, since the retweet path is an unknown feature, with too many variables and conflicting arguments.

The temporality of retweets has been the focus of some researchers, including Kwak et al. [35], who also used retweet trees as an illustration of the propagation pattern produced by Tweets. They found that, generally, half of all retweet action on a Tweet occurs *within an hour* of the Tweet being posted, and that by the end of the first day, 75% of all retweets of the Tweet will have been carried out. The authors also conducted an analysis on the elapsed time of a Tweet's travel between hops as it is retweeted. Although they observe a 'flatter' time initially, indicating that Tweets travelling over the first few hops are retweeted almost concurrently, they also found there to be a general incline in time taken for retweets to occur over the shorter path-lengths. After this point, the time taken becomes more 'noisy'.

An interesting notion that is not directly addressed in this thesis is that the time a particular Tweet is authored may have some effect on its propagation speed. Just as 'prime-time' television achieves higher audience ratings as it is at a time of the day when many people are at home and relaxing, Twitter may also exhibit a prime-time window in which its users are more active. For example, if a user posts a Tweet at a time when many of his/her followers are asleep, then the immediate audience size of the Tweet can be significantly reduced.

If there are fewer initial users viewing the Tweet, then the likelihood of retweet, as a function of this, is also reduced. This could have an effect on the perceived popularity of the Tweet, although since, by definition, there are fewer active users on Twitter at this time, then the number of Tweets sent during this period will be much smaller. Therefore, this is not taken into account during experimentation in later chapters.

3.4 Summary

In this chapter, a set of initial exploratory analyses have been undertaken into the behaviour of retweets and retweet activity in Twitter, the properties of retweet groups, the relationships between the propagation graph and the social graph, and briefly into the effects of time on Tweet dissemination.

The analyses were found to support and complement the findings of other research in the area, including the notions of message cascading [22] and the relationships of this to the interconnection of users on the social graph through communities [32]. Trees representing the retweet groups were found to grow in a variety of ways, from those illustrating long retweet chains, indicating a high level of inter-community dissemination, to shorter and wider trees, in which propagation can still be widespread but not as likely to disseminate to other communities.

User influence, in terms of an author's follower count, was observed as being an important factor in facilitating information spread, implying that popular users also produce popular information, since these users are more likely to achieve more retweets.

These inferences have helped to describe the multi-dimensional principles of retweet groups in terms of the features governing their spread over the social graph, and the quickness with which many users can be exposed to a Tweet. Although it is important to have an understanding of user psychology, and the thought processes behind the retweet decision, of most interest in this chapter is the analysis of the social structure.

3.5 Taking the Investigative Research Further

Twitter's social structure has been found to have a large effect on Tweet propagation, since it combines the features observed around user influence (in the naïve form of a user's follower count) with that of communities and sub-graphs of dense and sparse user interconnections.

In the following chapter, interests are focused on research questions **RQ3** and **RQ4**, in which the topological structure of user followships is studied through investigations into the flow of information between users arranged differently on the social graph. This research is conducted in order to develop a method to infer information interestingness, taking into account these information flow properties and user influence. It is clear that different Tweets can have a different level of *quality* in that Tweets that are retweeted have a greater chance of being interesting, but does the way in which the social structure of users is formed also have a 'quality' in terms of the propagation characteristics exhibited?

In the Background chapter, it was discussed that interestingness is a property of information that occurs when the information is different to what was expected. The interestingness inference method developed in the following chapter exploits this feature by comparing the expected popularity of Tweets to what was actually observed. From the research in this chapter, and from what is learned from the literature in the Background, it is clear that the retweet count of a Tweet is a good representation of Tweet popularity. Thus, understanding methods for producing an expected count is required for deducing interestingness, and this is addressed further as part of the network analyses in the following chapter.

Chapter 4

Analysis of Twitter's Social Structure

The social graph of Twitter describes how users are interconnected and fundamentally dictates information flow as Tweets are propagated through its structure. Research so far has focussed on the propagation of information through Twitter's social graph as a result of retweeting. In particular, this research provided an understanding of the patterns produced through retweets and how their properties relate to the users that relay the Tweets. Users with a higher follower count are more likely to have their Tweets retweeted, due to there being more users available to *see* each Tweet, and that some users can have their Tweets forwarded through many hops indeed, so that information may be passed between different communities of users.

In addition to the effects of user influence, several other factors also appear to influence an individual retweet decision of a given user for a particular Tweet. These include its properties, such as whether, or not, it contains a URL, whether it mentions a particular user, whether the user even has an opportunity to *view* the Tweet, and so on. These factors account for the individual user's retweet decision and the amalgamation of every user's retweet decision on the Tweet describes its overall retweetability, which determines how far it can propagate.

However, it is believed that the topology of the network, below the level of user influence and other factors, can play an important role in facilitating (or inhibiting) Tweet propagation by constraining the available retweet pathways between users and communities. Whilst retweet decisions based on Tweet features alone, such as its actual text or the contents of a document a URL it contains refers to, may imply a level of interest in the Tweet, clearly the influence of users can have a very large impact on how many retweets a certain Tweet receives. Thus, abstracting the concepts away from user influence may help in discovering methods for deducing which information is actually interesting.

Twitter's social structure has earlier been described as being built from users creating edges between themselves through the act of *following*. A followship defines the direction of travel of information from the friend to the follower, and this illustrates how users with many followers immediately have their Tweets made available to many more users before any retweeting even takes place. As more edges are constructed between users, the initial audience in terms of the number of users directly receiving the Tweets is increased, and, when the addition of retweets is considered, this effect is amplified. Although other intervening factors have been mentioned in earlier chapters, such as the notion of a user's network awareness and of user influence, the organisation of users on the graph and the differences in observed propagation pattern is a promising route for research towards uncovering the properties surrounding interestingness.

The research reported in the previous chapters help reinforce the disconnect between the social graph and the structure between users produced by retweet chains. That is, that the graph produced by retweet pathways is independent from the underlying social network, although research showed that there are strong ties, in some cases.

Due to these strong ties, it is felt that the social structure might have a large impact on how far retweets may be allowed to spread between users. It is known that interestingness of information can be derived from its distance from what was expected, and it has been explained that a Tweet's popularity can be derived from the number of times users have chosen to retweet it - the amalgamation of user *interest* in the Tweet.

Therefore, in this chapter, various social network structures are constructed in order to simulate retweet behaviour between users on Twitter. The behaviours are studied with the aim of researching the propagation patterns observed in different network structure types. Non-realistic and realistic graphs are built in order to highlight the low-level propagation characteristics in these networks and the similarities between more realistic simulated networks and Twitter's own social graph.

This research forms a basis for simulations of Tweets through Twitter's social graph as part of the development of a methodology for estimating Tweet interestingness based the distance of an observed Tweet popularity to *expected* popularity. If retweet counts can be easily derived from the network, then this might provide cues for generating this expected popularity.

The remainder of this chapter addresses question **RQ3** from the research questions identified in Section 1.3 and goes some way to answer question **RQ4**. More specifically, contributions include a study into the propagation characteristics exhibited by different graph structure types, and an analysis to explain the findings and illustrate the social graph's importance in retweet spread. From this research, an initial methodology for the non-semantic identification of interesting Tweets is built.

4.1 Propagation Patterns Exhibited by Different Graph Structures

Although it has been found that Tweets with particular properties may imply a certain quality that affects a user's retweet decision on the Tweet, of interest also is the effect of the potential presence of a graph 'quality', in that particular network structures may have an effect on how Tweets are spread.

In this section, to help in addressing this research area, simulations are carried out in three different network topologies - a path (or 'linear') network, a random network, and a scale-free network. In the experiments, individual user *decisions* are used as the bases for demonstrating retweet behaviour.

The simulation algorithm and ideas behind the model used for generating the simulated users' retweet decisions are adapted from the work carried out by Zhu et al. [66] and Peng et al. [46], who introduce methodologies for illustrating Tweet spread through a given network of users, and the simulations can be used to produce a retweet group for a given Tweet. From the analyses of the simulation experiments, of interest is whether, and how, changing the network structure does affect retweet propagation patterns, and whether a simulation can mimic Twitter's own behaviour in terms of retweet spread. If it is the case that the user structure does have a large impact on propagation, and since an individual retweet decision implies that user's interest in a Tweet, then this feature may be used as a basis for estimating interestingness.

Measuring retweet behaviour is carried out through studying the distribution of retweet group sizes that result from running the experiments, as is described in later sections.

4.1.1 Overview of the Simulation Algorithm

The algorithm covers the simulation of Tweet propagation through a given set of connected users by emulating retweet decisions of each user who receives the Tweet, as is described below. The retweet decision is made using a prediction based on a logistic regression classifier, which is discussed in more detail in Section 4.1.2.

Zhu et al. [66] developed a simulation algorithm which was found to be capable of accurately predicting retweet decisions using a logistic regression. These methods were modified and adapted to fit the purposes of the analyses in this section. The simulation is initialised with a graph of connected users, G_U, and a Tweet, t, which is introduced to the graph and retweeted between the users as described below. $G_U = (V, E)$ is the social graph of the vertices and edges representing the users who may receive t and the followships between them.

The method begins by initialising a set of users, S, to contain the followers of $t.\text{author}_O$. At each time interval, users in S form the set of users to have t or a member of $RT(t)$ currently on their home timelines and available to retweet. The procedure iterates over timesteps, at each generating a retweet probability on t, $P(u,t)$ for each $u \in S$. If $P(u,t)$ is greater than a pseudo randomly-generated $0 \leq r < 1$, then u creates a new retweet of t, which is added to $RT(t)$. u is then removed from S and the followers of u are added to S, since these users now also hold t and have the chance to make the retweet decision. u is now unable to retweet t again.

A threshold value, H, is used to emulate the notion of the Tweet 'decay' experienced when one uses a Twitter client or the web interface. The reasoning behind this is that as time goes by, more and more Tweets arrive onto the recipients' home timelines. This pushes the previous Tweets further down, whether they are interesting or not. Tweets may be ignored and not retweeted if the user has not viewed their home timeline for a while or if the user decides the Tweet is not of a sufficient quality to retweet it. If a Tweet is pushed down to the extent that is out of view, or out of the current page, then the chance of that user retweeting that Tweet is reduced. Thus, if a user is in S for more timestep iterations than specified by H, then the user is removed from S, meaning that it can no longer have the chance to retweet the Tweet. Users who have retweeted t, or are unable to do so (either by having previously retweeted it or by exceeding H) are prohibited from being (re-)added to S.

The algorithm terminates either when the timesteps thus far iterated exceed the maximum allowed, T, or when $S = \{\}$. This results in the retweet group, $RG(t)$, which comprises the final members of $RT(t)$ along with t. As in the previous chapter, $t.\text{count}_R = |RT(t)|$. Therefore, the additional necessary components to run the simulation are a user graph, an initial Tweet, and functionality for generating a retweet probability for each user who receives the Tweet.

The method is similar to that introduced by Zhu et al. [66] and Peng et al. [46], with the main differences in the selection of and reasoning behind feature selection, as explained over the coming sections.

Algorithm 1 Simulation of retweet decisions on t in a given graph, G_U

procedure SIMULATE(graph G_U, Tweet t)
  $RT(t) \leftarrow \{\}$
  $T \leftarrow$ total timesteps
  $H \leftarrow$ decay threshold ▷ Emulate t 'slipping down' timeline
  $S \leftarrow \left\{ u' \in V(G_U) : \exists \ \overleftarrow{t.\text{author}_O \quad u'} \in E(G_U) \right\}$
  u.TIME_HELD $\leftarrow 0 \quad \forall u \in V(G_U)$

  while $l < T$ and $|S| > 0$ **do**
    for all $u \in S$ **do**
      $P(u,t) \leftarrow$ retweet probability of u on t
      $r \leftarrow$ random number in range $[0,1)$
      if $P(u,t) > r$ **then**
        $r \leftarrow$ new retweet, where r.orig $= t$ & r.author$_R = u$
        $RT(t) \leftarrow RT(t) \cup \{r\}$
        $S \leftarrow S - \{u\}$
        $S \leftarrow S \cup \left\{ u' \in V(G_U) : \exists \overleftarrow{uu'} \in E(G_U) \right\}$
      else
        u.TIME_HELD $\leftarrow u$.TIME_HELD $+ 1$
        if u.TIME_HELD $> H$ **then**
          $S \leftarrow S - \{u\}$ ▷ u has held t for too long in timeline
        end if
      end if
    end for
    $l \leftarrow l + 1$
  end while
  Return $RT(t)$
end procedure

4.1.2 Generating a User's Retweet Probability

Calculating a value for $P(u,t)$ relies on building a model to represent t's features and its relationship with u. Zhu et al. [66] used a predictive model for retweet decisions based on a logistic regression, which was demonstrated to be capable of accurately predicting a user's retweet chance on a given Tweet. For the research in this chapter, a logistic regression model is also used, and was trained on a set of user, Tweet and context features in order to classify a likelihood on the binary decision of user u retweeting Tweet t. This probability is then compared to a randomly-generated r (as shown in Algorithm 1) in order to make the decision, such that if $P(u,t) = 1$ then u retweets t.

Machine Learning

Machine learning is the term given to the family of techniques that allow a program to make predictions for the outcome of unseen instances based on an observed and known history of occurrences. There are many types of machine learning classifiers that are suitable for different purposes, such as for predicting an expected outcome from a set of nominal categories, for predicting a value from a continuous range, or for predicting the *probability* of a binary outcome.

Most machine learning techniques involve the training of a predictive model, which contains the information on known outcomes for a set of features. The model is then used to estimate an unknown outcome, usually with a probability on the *confidence* of the classification, for new sets of instances.

For example, consider three attribute variables, A, B, and C, each of which can be equal to one of two nominal values; TRUE or FALSE. A particular machine learning algorithm trains a model based on its knowledge that;

- $A \leftarrow \text{TRUE}$, $B \leftarrow \text{FALSE} \Longrightarrow C \leftarrow \text{TRUE}$

- $A \leftarrow$ FALSE, $B \leftarrow$ FALSE $\Longrightarrow C \leftarrow$ FALSE

Although training of predictive models nearly always involves using more than two instances, the history of these example instances indicate that C is more strongly associated with A than with B. As more instances are added showing similar patterns, then the association becomes stronger, to the extent that the classifier will predict $C \leftarrow$ TRUE in instances where $A \leftarrow$ TRUE (and vice versa) with higher confidence.

In this case, A, B, and C are known as the 'features', and a set of such feature values forms the 'instance'. Once a trained model has been constructed, the machine learning algorithm will only be able to make predictions using instance features it has knowledge of. For example, if the example classifier was now given an instance containing a feature D, then it will not have knowledge of how changes in D will affect C's outcome.

If there is not a strong correlation between the features in a dataset, then the confidence of the classification of a particular feature will be weaker. Although this example has focussed on boolean (nominal) data types, many machine learning classifiers are able to work with features that are higher dimensional nominal values, continuous reals, and so on, and will apply weights to the different features based on their level of influence over other features in the instance.

Logistic Regression

Much of the research and experiments conducted in this chapter use logistic regression classification techniques in order to classify Tweets and other information. The Background chapter discussed other literature also using logistic regression for social network and retweet analysis [13, 66, 46, 43, 28].

The aim of logistic regression analysis is to find the closest model to map correlations between an outcome and a number of input variable features [29]. Examples of literature already discussed describe the use of logistic regression for 'predicting' the probability of a binary outcome in various tests. This type of modelling is appropriate for making predictions on the positive outcome of binary retweet decisions.

Once the model is trained from a set of features, $P(u, t)$ is obtained by testing an instance of features encapsulating the relationships between t and u (as explained in the following section) against the model in order to produce a probabilistic likelihood on these features producing a positive retweet decision.

4.1.3 Summary of Training Features

Zhu et al. [66] used a set of around 50 different features to train the logistic regression for the purposes of simulating retweet decisions, with the retweet outcome (TRUE or FALSE) being the predicted classification in each case. This set included Tweet-related features (such as content analysis, inclusion of URLs, etc.), and network and user features (followships, mentions, etc.). The authors achieve a precision and recall of 73.5% and 40.3% respectively in cross-validations of their trained model.

Since the network structures themselves, and the propagation *patterns*, are what are of interest in this section, the simulation is purposefully and significantly simplified by using far fewer features, yet ones which are features that have been shown to have a stronger influence on the retweet decision. By placing less of the retweet spread responsibility on the individual retweet decisions, and by abstracting them away more from the properties of the social graph, the importance of the structure can become more apparent. Zhu et al. [66] use many features relating to the semantics and content of Tweets, which are also not taken forward here in order to further accentuate the effects of the social structure on dissemination.

As such, each instance comprised the following four features associated with each Tweet, t, and where u is the user currently making the retweet decision, RETWEET. The default value for each feature is FALSE;

The URL feature has, in the literature, often been found as a large impacting feature

Feature	Data type	Description
FOLLOWS	{TRUE, FALSE}	TRUE if $u \in N^{+}(t.\text{author}_O)$
FOLLOWED	{TRUE, FALSE}	TRUE if $u \in N^{-}(t.\text{author}_O)$
MENTIONED	{TRUE, FALSE}	TRUE if u is mentioned in t.text
URL	{TRUE, FALSE}	TRUE if `http://` or `https://` in t.text
RETWEET	{TRUE, FALSE}	TRUE if $u \in RT(t)$

Table 4.1: Training features for the logistic regression for simulating retweet decisions.

on retweets in Twitter, especially by Alonso et al. [4], who use it as their basis for determining and identifying interesting Tweets.

4.1.4 Training the Model

In order to train the logistic regression model, data was required from Twitter so that the sets of feature instances could be built.

Data collection for these experiments again utilised Twitter's REST API, which was queried between March and June 2012 to collect a set of around 18,000 Tweets. The data was collected as part of a random crawl through the social graph, starting at one user and choosing a random follower of the current user to use as the next step. At each stage, information on the current user and on a set of that user's recent Tweets were collected. Tweets that had 0 retweets were collected in order to provide features presenting a negative case when training the regression model and to ensure that there were instances where the RETWEET feature could be FALSE. In this dataset there are around 2,600 Tweets (15%) that had been retweeted at least once.

In cases where the collected Tweet had been retweeted, further calls were made to the API to determine the relationships between the authors of the retweets and the original Tweet's author in order to satisfy the required FOLLOWS and FOLLOWED features.

Where the collected Tweet had not been retweeted, there are no retweeting users to examine the relationships between. In these cases, further Tweets were retrieved for the user in order to find their retweet rate in terms of the ratio of retweets to Tweets on their user timeline and an analysis of the relationship between these and the original authors. This was used in conjunction with the user's follower and friend count to determine a probability of the 'faux' followships. As mentioned, the accuracy of the retweet counts obtained through the simulations is not particularly important; of interest is the propagation patterns observed over the graph structures.

After storage, the regression model was trained using features (see Table 4.1) extracted from the raw data, with the outcome being the binary retweet decision. The algorithm could then use the model to generate the required retweet probability, $P(u, t)$, by classifying each user's RETWEET decision outcome at each stage.

4.1.5 Running the Simulations

The logistic regression model, having been trained with the features collected for the training dataset, could then be tested against newly-generated Tweet features in order to output the probability, $P(u, t)$, indicating u's retweet decision likelihood on t. The simulation method only produces probabilities for those $u \in V(G_U)$ that actually receive t and have a chance to retweet it. Since some of the features used rely on the relationships of the current u to t.author$_O$, not every user in the graph will have the same value for $P(u, t)$.

For each simulation experiment, a network of users was generated according to a structured model, as described in the next section. A Tweet object was then constructed and contained information on whether or not it contained a URL and if it mentioned one of the users in the generated network.

In each network analysis, the same set of Tweets was used. This set comprised Tweets generated from many feature combinations. Various parameters - such as H, n (the size of G_U to be generated), and any weightings on the decision probability prediction generator - were altered in the simulations in order to affect the visibility of any correlations in the propagation patterns produced by the different structure types. As such, the *volume* of retweet counts produced are not comparable across the structure analyses, but the correlation patterns are.

4.1.6 Network Analyses

In this section, three network structures are assessed in terms of the differences in the patterns of propagation each permits. Each generated graph is *directed* in order to illustrate the followships between the user nodes, and to support the use of the FOLLOWS and FOLLOWED features required in the decision probability calculation.

Path Network

The first assessment involved illustrating the propagation pattern observed in the most basic social network structure; a path network. Albeit non-realistic in practice, these graphs represent the fundamental structure of a connected 'world' of users.

A linear directional path network consists of the graph of users, G_U, of size n;

$$n = |V(G_U)| = |E(G_U)| + 1$$

where;

$$\exists \quad \overleftarrow{u_i \quad u_{i+1}} \in E(G_U) \quad \forall \quad 1 \leq i < n$$

As a result, each $u \in V(G_U)$ has precisely one follower and one friend, except the users u_n and u_1 respectively. n is the only parameter necessary in the construction of this user graph.

In this graph, the size of the retweet group is, by definition, equal to the depth of penetration, as there is only one path (or retweet chain) available for propagation to occur along. As such, in each case, the retweet tree representing a resultant retweet

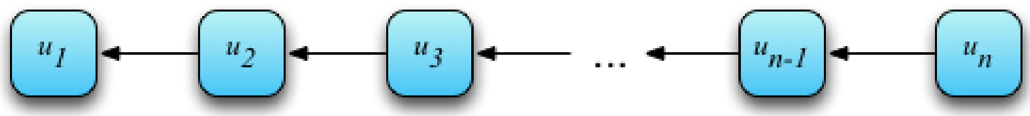


Figure 4.1: Example of a path network.

group formed in this type of network will have the same structure as the graph itself, with a size dependent on the collective retweet decisions of the users.

Since each internal user has only one follower, the likelihood of each progressive user in the graph being able to view the Tweet in order to make the retweet decision reduces, and thus the retweet count is much more likely to tail off sooner than in graphs with more propagation avenues. This is also due to the fact that each retweet can only reach an audience of size 1 at each time step, and thus the 'survival' of the Tweet cannot rely on a summation of many users' retweet decisions. The actual retweet *decision* is not affected by a user's position in terms of 'distance' from the source except through the effect of the regression features.

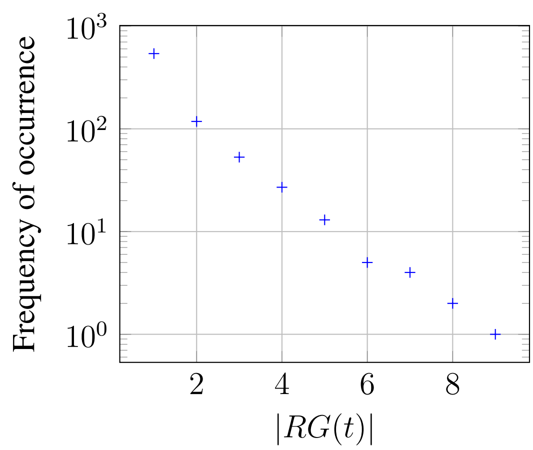


Figure 4.2: Frequency distribution of retweet group sizes in path network simulations.

The likelihood of a particular user achieving the opportunity to receive the Tweet, in order to then retweet it, becomes the product of the probability function the further it travels through the graph, in which user u_i requires each user from u_1 to u_{i-1} to first receive it and then make a positive retweet decision. For example, if each user has probability p of retweeting the Tweet, then each user's chance of receiving the Tweet is p^{i-1}, where i is the position of the user in the graph. A user can only make a retweet decision on t once it has been received.

The path network analyses involved 50 repeat simulations of the set of Tweets on a graph of size $n = 1000$. A timeline threshold (H) of 30 was used to represent the maximum time a Tweet is permitted on a user's timeline before it is no longer retweetable, for reasons discussed earlier.

As might be expected, the frequency distribution of retweet group sizes in Figure 4.2 shows a half-life type behaviour demonstrating the logarithmic pattern with many small retweet groups followed by a series of exponentially smaller groups. This user structure illustrates well how some users that might find the Tweet interesting, and who may then decide to retweet it, do not even get the chance to view it in order to *make* that decision. Although this is accentuated in this structure, the same principle applies to any non-complete social graph, and demonstrates how the way in which users are connected can have a large impact on the overall retweetability of a particular Tweet.

Random Network

The random network was the next user structure to be analysed. Although it is certainly more similar to a real-life social graph than a path network, it is much more basic and uniform and does not consider user communities and clusters or different levels of influence in users in terms of differences in follower and friend counts.

A random social network is in this case based on the Erdős-Rényi model [20] and defined as a graph, G_U, where $n = |V(G_U)|$, consists of each user, $u \in V(G_U)$, having the connection probability P_c of following each other $u_i \in V(G_U) \forall 0 \leq i \leq n$ and where $u_i \neq u$. Thus, as P_c is increased, then so does the likelihood of u following a u_i, causing the network to have a greater overall edge density. In general, therefore,

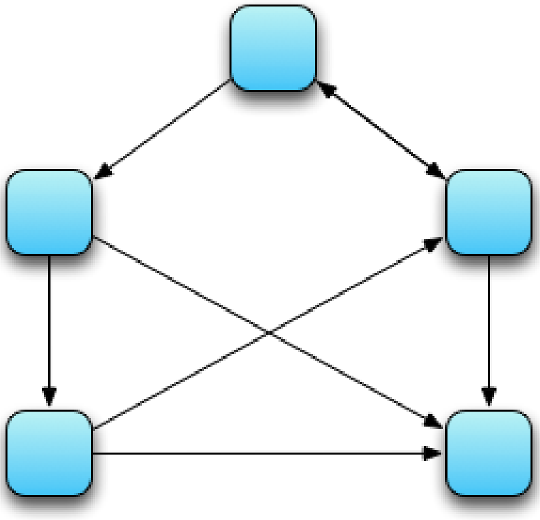


Figure 4.3: Example of a random network where $n = 5$ and $P_c = 0.5$.

the average number of followers and friends of a user is proportional to $P_c.n$. The only parameters needed for constructing such a graph are n and P_c.

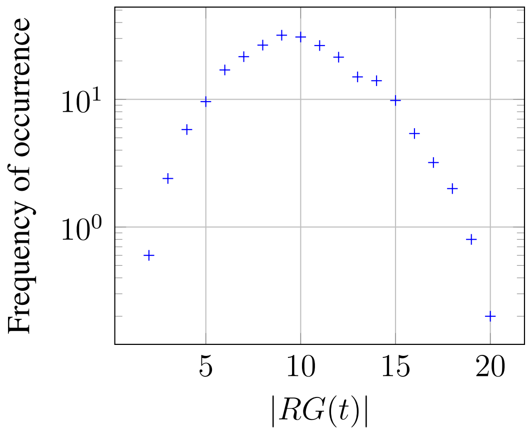


Figure 4.4: Frequency distribution of retweet group sizes in random network simulations.

As with the path network analysis, a graph size of $n = 1000$ was used under 50 simulations with the value of $H = 30$. The distribution in Figure 4.4 was generated using a value of $P_c = 0.01$, meaning that, on average, each user had 10 followers and 10 friends and was felt to be representative given the size of the entire graph.

The frequency distribution in Figure 4.4 demonstrates a very large proportion of mid-range values for $|RG(t)|$, indicating that Tweets tend to have a consistent spread amongst the network, as might be expected. There are few smaller groups since there are no users that have disproportionately smaller spheres of influence, and each user has many incoming edges and a similar number of outgoing edges. This explains the smaller number of lower-range retweet group sizes observed in the distribution. However, as in any distribution so far examined, the number of larger retweet groups must eventually tail off due to the natural eventual reduction in positive retweet decisions being successively made as propagation chains increase in length.

Scale-Free Network

The final network structure examined in this section is the scale-free network. Also known as 'small world', scale-free graphs are generally known to be representative of the general structure of 'real-life' and online social networks [40] and, indeed, they are also used to describe the interconnections of real-world properties, such as friendship groups and food webs [12, 26]. Essentially, scale-free networks dictate that there are a small number of nodes with a high degree and many nodes with a low degree, and are usually generated through some form of preferential attachment algorithm. Thus, this type of network has support for the consideration of user communities and influential users in terms of those demonstrating a disproportionately large follower count. The other user structures studied do not have the scope for emulating this property of inconsistent interconnection between the user nodes.

Scale-free networks are constructed such that the distribution of the degree of the graph's nodes follow a power-law in that the distribution of the number of vertex edges across the graph is logarithmic. For these analyses, NetworkX[1], a Python graph and networking package, was used to generate directed scale-free graphs of users, based on a graph size, n, and other arguments, including δ-out, as the graph construction

[1] http://networkx.lanl.gov

parameters.

For the scale-free analysis, the same graph construction and simulation parameters were used as in the previous analyses. δ-out represents the bias for a node's selection for out-degree from the other available nodes, and was set to a value of 0.7 to improve the clarity of the distribution result.

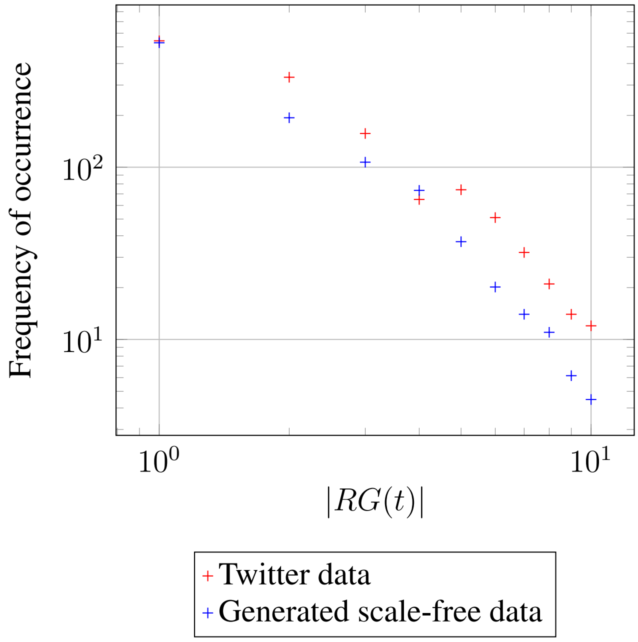


Figure 4.5: Comparison of retweet group size distributions from scale-free graph simulations and data from Twitter's own social graph.

From simulations of the algorithm through these scale-free networks, a logarithmic trend is observed similar to that demonstrated from the 'real' Twitter data analysed in the previous chapter and published by Webberley et al. [57], and the similarities in the distribution pattern is illustrated by Figure 4.5.

4.1.7 General Comparison of Propagation Characteristics across Different Graph Structures

In this section, three different network structures have been compared, and whilst the path network is very unrealistic as a representation of a social network, the differences in propagation behaviour presented by each do show how the interconnection of users on the graph can have a large effect on the spread of a Tweet. A small set of features to govern retweet features were used in order to accentuate the difference made by the user structures themselves.

This has demonstrated that, in addition to the processes behind a user's individual retweet decision, the eventual spread of a Tweet also depends on how the original author's local network is arranged. Thus, the retweet decision of each involved user along with the available information pathways provided by the underlying social structure both contribute to the overall retweetability of a Tweet.

If there are many edges in the network, such as in the case of the random network, then there are many more routes for propagation to occur down due to the relatively large in- and out-degree of each user node on the graph. This increases the number of users who end up receiving the Tweet and then have the chance to make a retweet decision. This resulted in there being a larger distribution of larger retweet group sizes than smaller ones, before naturally diminishing again.

Despite this high throughput of retweets, which provides a high level of information *recall* for the users, the random graph structure is likely to demonstrate a low *precision* in terms of the interestingness of the received Tweets. This is due to the large number of users having the opportunity to retweet the Tweet, increasing the chance that 'noisy' information will be filtered through. As illustrated in Figure 4.6, it is generally the case that if a person follows more users, they will receive exponentially more noise due to its prevalence over the interesting information (a decrease in precision), yet will likely receive more of the interesting information that is available (an increase in recall).

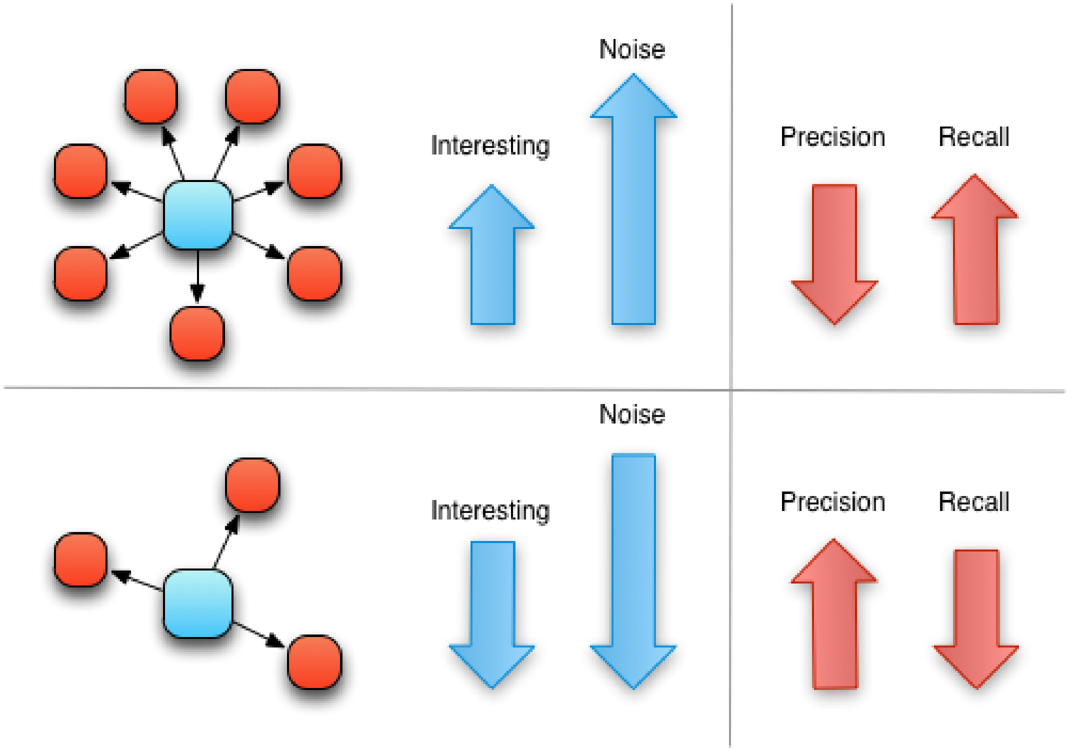


Figure 4.6: Comparing the effects of followship decisions on precision & recall.

The path network demonstrated the opposite effect in that its nodes can only follow a maximum of one other node. This demonstrated very poor propagation, and required its simulation parameters to be altered to facilitate retweet behaviour more significantly than in the other graph structure analyses in order to produce any observable pattern. The results showed that propagation down a single allowed chain cannot be an effective way to spread Tweets, as it required each user in the chain to retweet it so that the successive users can have a chance to view it.

Similarly, in Twitter, users who follow very few others are likely to be more selective about who they follow. They will therefore achieve a greater precision in terms of the interestingness of information received, but the recall will be much smaller (Figure 4.6). Generally, it is impossible to achieve a perfect precision and recall as it is likely that there will always be interesting information not being received, and any noise that is received at all will reduce the precision.

Whilst the scale-free network does not have the same general propagation throughput as the random network, it does demonstrate retweet patterns similar to those observed in data from Twitter's own social graph. This complements the findings of Mislove et al. [40] and Hein et al. [26] in terms of online social networks emulating real-life social networks having scale-free properties. This type of structure supports areas of the graph with denser communities, as is shown to exist by Java et al. [32], and have the potential for facilitating very large numbers of retweets if influential users are involved, but illustrate how Tweets 'travelling' through less dense areas (and less-influential users) will not be as demonstrably popular.

4.2 Using the Social Graph as a Method for Inferring Interestingness

The graph analyses in the previous section have demonstrated a method for generating a $RG(t)$ for a given Tweet, t. Since $t.\text{count}_R = |RG(t)| - 1$, then the same simulation algorithm can be used to estimate a retweet count for a given Tweet. The analyses conducted in the previous section relied on modelling the retweet *decisions* made by the users, which individually account for that particular user's interest in that Tweet. Although it has been previously discussed that the overall retweet count cannot realistically be used alone for determining the level of interest in a Tweet, it is clear that interestingness of a Tweet is certainly based on a *function* of the culmination of positive retweet decisions being made on the Tweet.

This notion is based on the idea that if a specific Tweet is more popular than the model predicts, then there is something about that Tweet that makes it more *interesting* than similar Tweets that are less popular, such as a piece of breaking news or a link to a controversial article. For example, consider the case of two Tweets, written by the same author, and both containing the same instances of feature values in that they both contain a URL and mentions a user. If one of these Tweets achieves significantly more retweets than the other, then there must be some non-structural feature of the more popular Tweet that makes it stand out to the audience, and thus allows it to be perceived as more *interesting*. This is because the features taken into account are static, and do not address any semantics of the actual content of the Tweet.

Similarly, if most Tweets of a user achieve between one and two retweets, then the expected retweet count for this user's future Tweets is likely to be similar. If, however, the author posts a Tweet which achieves an observed total of 10 retweets, then this is more popular than expected. If a Tweet achieves one or zero retweets, then this is as expected or less than expected, and is therefore not interesting.

Proposition 4.1

> The interestingness of a Tweet is a function of its expected and observed popularity. In particular, that a Tweet should be labelled as interesting if it is more popular than it was expected to be.

This form of analysis is akin to anomaly detection and, in particular, the Helmholtz principle, which is defined by Balinsky et al. [9] to be the deviation of interesting events from the randomness of the non-interesting events surrounding them. In the context of Proposition 4.1, exceptionally popular (or unpopular) Tweets can be represented by the interesting events, and the *norm* of expected Tweet popularity is analogous to the randomness of non-interesting events.

Figure 4.7 helps demonstrate this behaviour in the context of Proposition 4.1, in which there is a 'baseline' of expected Tweet popularity given by the properties of the Tweet and its environment. The exceptional, or anomalous Tweets, are the ones that are significantly different from the baseline; those that are much higher than this baseline are labelled as interesting since they are more popular than their features imply, and vice-versa.

As such, a method is proposed based on the following two criteria;

- $t.\text{count}_R > e(t) \implies t$ is interesting

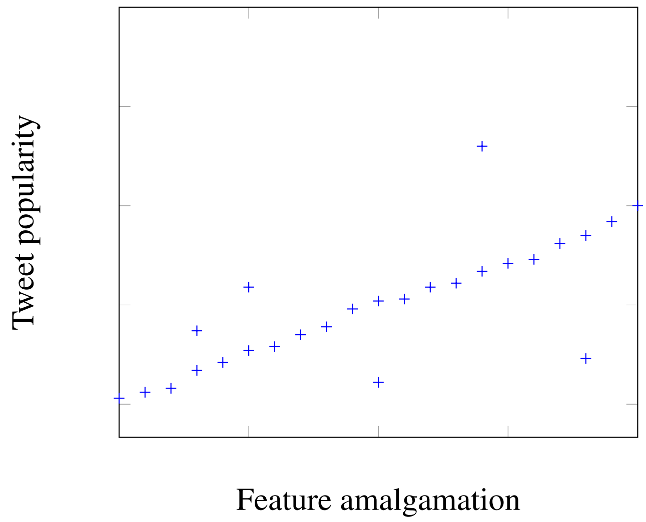


Figure 4.7: Conceptual example illustration of Tweet popularity as a function of their properties.

- t.count$_R \leq e(t) \implies t$ is non-interesting

where $e(t)$ is the expected retweet count of t.

Although it was found in the previous chapter and in other relevant literature that pseudo-generated scale-free networks can be representative of Twitter's own social structure, a user's actual own local social network would more accurately portray the links between the users surrounding the original author of a Tweet. By constructing a network based on a user's own local network, then the method would effectively be simulating the Tweets' propagation through the edges representing the followships of the real and appropriate users in Twitter's social graph.

Thus, in the simulations, the user in question is t.author$_O$, t is the Tweet to be simulated and initially $S = N^+(t.\text{author}_O)$. At each timestep, each user in S would have the opportunity to retweet t, and therefore, by running the simulation multiple times, an estimation of $e(t)$ can be obtained.
In particular, the method follows these steps;

1. Select a user, u, from Twitter to be t.author$_O$

2. Collect u's local follower network
3. Collect a set of u's recent Tweets
4. Construct a network based on the users and edges of the collected network
5. Simulate the collected Tweets through the constructed network using the simulation algorithm.

This procedure would provide an estimated retweet group size for each Tweet, which could then be compared to the actual observed retweet count of the Tweet on Twitter to help towards deducing the interestingness.

4.2.1 Data Collection

In order to simulate Tweets through Twitter's own social graph, data on its users and edges is required so that a copy of the graph can be built locally. This is necessary for the users' retweet decisions to be modelled and to keep track of which users are able to receive the Tweets.

Ideally, the actual social graph would be used, but due to the scaling properties encountered in a breadth-first traversal of Twitter's social graph, it became infeasible to collect a user's local network containing users more than two edge 'hops' away from the source user under the rate limitations of Twitter's REST API. As previously described, v1 allowed 350 calls to the API each hour for each authenticated Twitter account. One call, for example, was required to obtain a list of up to 5,000 user IDs representing the followers of a particular user - the users one hop from the source user. An additional call would then be required to collect each of these user's own followers in order to provide the 2-hop representation of the local network from the source user.

For a user with a follower count of 700, a total of 701 API calls would be required to collect the user's local network within two hops - the one to retrieve the source user's immediate followers, and then one further call for each of the 700 followers. This would take over two hours of collection, and to collect the third hop would require another exponential number of API calls. If each of the 700 followers of the source user has, on average, 200 followers, then this would require a further $700 \times 200 = 140,000$ API calls, which, in total, equates to over 402 hours of data collection time. Although some follower overlap is likely to be present among the users two hops away, when one considers that this is simply the time taken to collect the local network for *one* user, then it becomes clear that this must still be an impractical approach. It would, however, be feasible for Twitter itself to use this data for these purposes.

Observation 3.1 states that the vast majority of retweets do actually occur *within* two hops of the source user, in that the most significant number of retweet groups analysed had a maximum path-length of less than three. In addition, as mentioned, online social networks are ‘closer’ than real-life social networks, and was found to have a value of around four degrees of separation in Facebook. These points help to justify the decision made to classify a user’s local network as those users and edges existing within two hops from the source user.

In June 2012, the Twitter REST API was used in order to conduct a random walk through Twitter’s social graph. Starting by selecting an initial user, an edge expressing the followship of a random follower was chosen in order to select the next user. This continued for each of the selected users in turn and, for each user selected, the most recent 300 Tweets and surrounding information was collected along with that user’s local follower network within two hops. The friend network (i.e. the outward edges from each user) was ignored, as only the directional outward flow of information from the source user was useful in this experiment. If, at any stage, the currently selected user did not have any followers, the collection algorithm backtraced to the previous user and another follower was selected instead. The crawler continued until the rate limit for the current request window was met, at which time the current data state was stored, and then waited until the rate-limit was reset before continuing.

The data collection resulted in a set of 33 Twitter users, each with a full 2-hop local network collected and a set of up to 300 Tweets. In total, around 10,000 Tweets were stored as a result of the crawl to be used in the simulations. It was decided that the previously trained regression model would be re-used as part of the retweet decision engine in this experiment also, and so no further training data was required to be collected. From the Tweets collected, the URL and MENTIONED features could easily be identified, and the two user features could be extracted under the same process as the one used in the network simulations in the previous section.

For each Tweet collected, a simulation could now be run in order to provide an expected retweet count for that particular Tweet. By comparing this value to the actual retweet count expressed by the Tweet, which is returned as part of the standard Twitter API call, an indication of whether or not the Tweet is interesting could be obtained.

4.2.2 Validating the Accuracy of Inference Results

In order to test the validity of the results, it was necessary to use human assessment on each of the evaluated Tweets to check for agreement between the interestingness inferences made by the algorithm and by humans. Although interestingness is a subjective notion, the validations were carried out in such a way to emphasise a *global* level of interest in terms of the general separation between noisy and un-noisy Tweets.

Crowdsourcing Validations

Crowdsourcing is a technique that has grown in popularity over many domains in recent years, including in the media, reviews services, sensor networks, and others. Essentially, crowdsourcing involves the use of many people (or, in some cases, devices) providing input or results on a given task. Services such as Google Maps[2], TripAd-

[2]http://maps.google.com

visor[3], and Stack Overflow[4] respectively use crowdsourcing for obtaining information (such as photos) on geographic locations, service reviews, and programming assistance. Its use means that the crowdsourcers can easily receive lots of input with very little additional work, since the load is spread amongst many people.

Crowdsourcing has also proved to be a useful asset in research as it facilitates the harvesting of many inputs, from diverse opinions and views, much more quickly than without it, and it is a useful tool for validating data. Many crowdsourcing services are active on the Internet to cater for different use-cases.

Amazon's Mechanical Turk[5] allows crowdsourcers to create small jobs (known as 'microtasks') to be completed by crowdsourcees, known as Mechanical Turk Workers (MTWs), who have an account on the website. The crowdsourcer describes the particular microtask in terms of what is expected of the MTWs and also determines the amount paid for the task. A single microtask completed by a particular MTW is known as a 'judgment', and MTWs are paid for each judgment he/she completes. The crowdsourcer can define certain criteria on the microtasks, such as allowing each MTW to only complete one microtask.

Due to Mechanical Turk's availability only to US credit card holders at this time, a third-party service, CrowdFlower[6], was used instead to submit the microtasks to Amazon's system in order to be completed by the workers.

Aims of the Validations

The purpose of the use of crowdsourcing was for evaluating the effectiveness of the interestingness inferences made through the comparison between an expected and observed popularity of a given Tweet. Of particular concern was the correlation between

[3]http://tripadvisor.co.uk
[4]http://stackoverflow.com
[5]http://mturk.com
[6]http://crowdflower.com

those Tweets that the algorithm denoted as interesting and the Tweets that humans found interesting.

Since the accuracy of the various components of the technique could not be known until they were properly validated, it was decided that the crowdsourcing would initially be run as a pilot test in order to identify the presence of any correlations. If this was sufficiently successful, then a further and more rigorous test would take place, involving more workers.

Constructing the Questions

The microtasks presented to the MTWs each consisted of a question containing five Tweets that were selected at random from the experimental set. Each question asked the MTWs to select which one of the five Tweets was the most interesting and which one was the least interesting. The validations were set up so that each of the questions was assessed by at least three different MTWs.

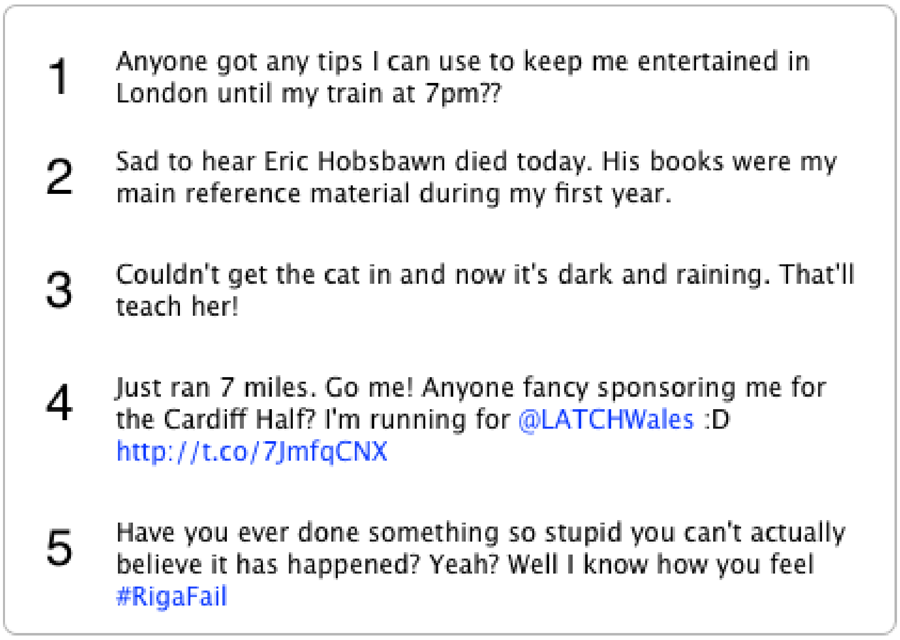


Figure 4.8: Example question for the MTWs: "Select the most interesting Tweet and the least interesting Tweet from the five shown".

Figure 4.8 shows an example of a question asked of the workers. It demonstrates also how the MTWs were not provided with any additional context, such as the author's username or the post date and time, for any of the Tweets to be evaluated.

Although Tweet selection was random, those whose content starts with a user's "@" username (i.e. '@-replies') were excluded, since these Tweets typically form part of a conversation between a small number of users and are unlikely to convey any interest to those not directly involved. The final validation set consisted of a total of around 4,300 Tweets to be assessed in the questions, and MTWs were encouraged to follow links to any websites or media included in the Tweets' contents as part of their evaluations. For each question answered, MTWs were paid $0.03.

Inference Performance Validation Results

The Mechanical Turk validation took place between the 7th and 11th December 2012. Table 4.2 provides further information about the validation tests. An assessment is defined as a MTW answering a particular question with his/her opinion on the most interesting and least interesting Tweet. Since there were 856 questions, there were 2568 assessments in total.

Information key	Value
Total assessments by MTWs	2568
Number of distinct questions	856
Number of unique MTWs	177
Number of unique Tweets	4280
Num. questions with $\geq \frac{2}{3}$ agreement on most interesting	510
Num. questions with $\geq \frac{2}{3}$ agreement on least interesting	493
Amount paid to MTWs per assessment	$0.03
Total amount paid	$77.04

Table 4.2: Information on the Mechanical Turk validation results.

A confident answer is defined as the case in which at least two of the three MTWs answering a particular question agree on the most interesting and least interesting of the Tweets. It is assumed that if at least two people agree on a piece of content being interesting, then this provides further strength to the individual assessments, and any questions that were not confident were excluded from the following validation analysis.

Through the retweet simulations and algorithm for each Tweet, an 86% accuracy was achieved in terms of correctly predicting the actual retweet count - the cases where the expected retweet count is equal to the observed retweet count. In around 30% of cases, a Tweet that was determined to be interesting through the methodologies described in this chapter was also verified as interesting by the agreeing MTWs.

This is a relatively low precision, and while it does mean that the method was able to correctly identify an interesting Tweet from a set of five in 30% of cases and the random performance of selecting an interesting Tweet could not reach this accuracy, it is not a strong enough result to describe the method as being suitable in the general inference of interesting Tweets. Despite this, Yang et al. [61] conducted a similar study aimed towards predicting whether or not a Tweet would be retweeted, and achieved a very close precision of 29%. This represented a better performance than two baselines the same authors assessed against, yet the method being validated in this thesis is currently unsuitable for inferring interestingness for a wide range of Tweets, as explained further in the following section.

As such, further investigation would be required to address the method with the aim of improving the method's applicability and performance.

4.2.3 Improving The Interestingness Inference Performance

Proposition 4.1 is still considered to be a viable way of addressing the problem of identifying interesting information for the reasons discussed earlier. However, a more convenient and accurate method is clearly required for acquiring the *expected* retweet value.

The issue with the current method is two-fold; a large volume of data is required in order to reconstruct the Tweets' authors' local networks in which to simulate the Tweets, which leads to the second problem of only being able to simulate Tweets from authors in sparser local networks. Under the current scheme, only users with a small enough local network (i.e. users that have lower follower counts) can realistically be evaluated, due to the collection criteria discussed previously, meaning that the methodology cannot be used in the general case. Although a high accuracy was achieved in predicting the *correct* retweet counts for the Tweets assessed in this section, most of these Tweets only actually had an observed retweet count of 0 or 1. This is the by-product of the previous issue in that only users with fewer followers could have their Tweets simulated, and these users will therefore typically receive few retweets per Tweet. Ideally, the methodology should have the capabilities to be applied to any type of user and any Tweet on Twitter.

Additionally, this method alone does not make efforts towards evaluating the *level* of Tweet interestingness. Instead of the binary interesting / non-interesting decision, it would be more useful to award each Tweet a score denoting the estimated interestingness of the Tweet. The further importance and usefulness of this is explained in further detail in the following chapter.

4.3 Chapter Summary

In this chapter, an analysis of propagation through differing structures of user connections on social graphs has been conducted. From this, a potential methodology for inferring interestingness of Tweets has emerged, which, despite being negatively impacted by various factors in its current form, shows promise as a suitable technique towards assisting in this task.

Question **RQ3** from Section 1.3 has been addressed in order to exhibit the differences in the propagation characteristics permitted by each graph structure type. Since this has been shown to be an important factor, the features will be taken forward as the research continues in the following chapter. Question **RQ4** has been partially answered, but the initial analyses did not show the derived method to be able to accurately infer Tweet interestingness. As such, more work is required to address the question more fully.

4.3.1 Network Structure Analysis

A logistic regression model was built as part of a simulation algorithm in order to anaylse the propagation characteristics of three different network structures; a path network, a random network, and a scale-free network.

Although the actual retweet counts of simulated Tweets in each network structure are not comparable due to the parameter alterations that were required in order to amplify visible results, the actual *pattern* of propagation in terms of the distribution of retweet group sizes was found to be different in each structure and for differing reasons. In addition, the scale-free network was found to express a similar pattern to that observed from the data on retweet group sizes discussed in the previous chapter.

4.3.2 Interestingness Inference Methodology

The model and techniques behind the network structure analyses were then applied to the goal of detecting the interestingness of Tweets based on the comparison of the expected retweet value, generated through the same algorithm used to simulate Tweets in the network analyses, and the actual observed retweet count of the Tweet.

Validating the methodology showed that the technique is not particularly useful in determining interesting information, and its other drawbacks, such as its application only realistically being available to Tweets from non-influential users, mean that the technique cannot be used in the general case. Further to this, the data collection required is not suitable for quick evaluations and may not remain accurate over time even after collection due to the continuous changing nature of the edges in online social networks as users create and destroy followships. This is particularly impactful in this case as there are many users involved even in a user's 2-hop local network.

In the next chapter, the methodology for generating expected retweet counts is adapted with the aim of improving its validation performance, the ease of preparation through data collection, and of addressing the methodology's current restrictions on the types of users it is suitable for. It is known from work in this chapter that the network structure plays an important role in information propagation, so this and more environmental features are taken forward as part of the improvements. Using the network features as a primary indicator for the expected retweet count estimation further distances the approach from semantic methods, and allows for classifying information on a global interestingness level distinct from individual interests and relevance.

Chapter 5

Inferring Interestingness

A method has been introduced, based on Proposition 4.1, for identifying interesting Tweets. However, this method presents a range of various shortcomings and was found to be un-usable for a large proportion of Tweets. In this chapter, the methodology is modified with the aim of improving its performance and increasing the range of use-cases considered. Since the social structure was found to play an important role in propagation, improvements are centred on including network and user features.

A modification representing a larger contribution is made in order to provide an indication of *how* interesting a piece of information is estimated to be, and more about this particular component is discussed in later sections. Motivation for this is based around wanting to rank Tweets by order of relative interestingness and to highlight the Tweets that may be brought forward to receive further attention.

The proposed methodologies build upon the previously-identified differences between a Tweet's raw popularity, as indicated by its retweet count, and how interesting the Tweet actually is to those who read it. It has been shown that making retweet predictions against models trained with a large number of features can be accurate [66], but in this work the focus is applied more to the Tweets' contents and properties beyond their static features. That is, that when comparing Tweet popularity, then there may be some content, either within the Tweet itself or perhaps in a resource indicated by a URL contained in the Tweet, that makes the Tweet stand out more to its recipients and to cause its readers to be affectively stimulated.

Of course, this brings about the notion of information *relevance*, and the fact that the same Tweet could be very boring or irrelevant to one user, and very interesting to another. In this work focus is applied to *global* (or 'average') interest, where interestingness inferences are made for the general case. It is considered that Tweets that are retweeted more than expected within their authors' local networks, relative to the usual retweet count of the authors' other Tweets, are also likely to be of interest to a wider audience, especially since they are now more likely to penetrate through the social graph enough to be received by users in different communities.

User influence plays a large role in the exhibited difference between popularity and interestingness. The Background chapter illustrated the example of Justin Bieber, whose account, `@justinbieber`, is one of the most 'influential' on Twitter, with nearly 50 million followers at the time of writing. His Tweets receive an average of around 50-120 thousand retweets per Tweet, and they rarely receive fewer than 40,000 retweets. Since an average Twitter user would generally attract a maximum of a few hundred followers, and would normally receive very few, if any, retweets per Tweet. A particularly interesting Tweet from such a user may be retweeted, for example, between 5-20 times. It is apparent that, in the general case, an uninteresting Tweet from an influential user may receive 50,000 retweets, and an exceptionally interesting Tweet from a less-influential user may be retweeted 30 times. Therefore, user influence dictates that this value cannot alone be indicative of Tweet interest.

However, since interestingness *does* have an effect on an a user's individual retweet decision on a particular Tweet, this absolute retweet count can be used as part of the method for generating an interestingness *score* for that Tweet.

The research so far has culminated in the development of a method, based from the Background chapter's research into 'interestingness' as a property versus 'interesting' as an adjective, for inferring whether or not a Tweet is generically interesting. As described, the method relies on the generation of an *expected* retweet count, which is then compared to an observed count to determine its global interestingness. The estimated expected retweet count generator based directly from local network simulations had many shortcomings in performance, accuracy, and apliccability.

As such, the focus of the work in this chapter is that of adapting the inference methodology in order to develop a technique for accurately *quantifying* the interestingness of Tweets. This is concerning universal relevance in terms of highlighting interesting Tweets from the noise. In particular, there are two main improvements over the previous methodology to be made;

- Improve the method for generating the *expected* retweet count of a Tweet (in both accuracy and in range of application);
- Expand the binary retweet interesting inference into a more useful scale in order to support the *ranking* of information by interest.

In summary, the research reported in this chapter involves addressing the method for generating expected retweet counts for Tweets in order to improve interestingness inferences. The final question (**RQ4**) from the hypotheses in Section 1.3 is answered in order to show that Tweet global interestingness can be inferred non-semantically with some degree of accuracy. Contributions made as part of this research include an analysis into the performance of machine learning classifiers for the purpose of social network analysis, and a thorough analysis into the performance of the methodology is made for the purposes of demonstrating its relative advantages and disadvantages, and how these relate to the social graph analyses considered earlier in this thesis.

5.1 Interestingness through Tweet Scoring

A scoring scheme is introduced in order to address the notion of interestingness quantification, allowing certain interesting Tweets to be ranked as 'more interesting' than other interesting Tweets. This, in itself, is an improvement over the previous method, which allowed only for Tweets to be labelled as 'interesting' or 'non-interesting'.

As an enhancement to Proposition 4.1's method of comparing the observed to expected retweet counts, the new scoring technique is based now on the *distance* between the two counts. The general idea and potential use-case for this is that if a score is known for a set of Tweets, then these can be used as a basis for ordering information as part of information retrieval or an information delivery system, where Tweets can be displayed to users in a more useful way than simply chronologically. In this way, interesting Tweets could be brought forward to users who don't follow the source user or a retweeter and thus deliver information to an interested user, yet without him or her having to know about it first.

Proposition 5.1

> If the positive difference between a Tweet's observed and predicted popularity is proportionately greater than those attributes in a different Tweet, then the first Tweet is proportionately *more* interesting than the second.

Essentially, Proposition 5.1 stems from the following scenario. Consider two Tweets, A and B, which have the following properties;

- $e(A) = 3000$ and $A.\text{count}_R = 3010$
- $e(B) = 5$ and $B.\text{count}_R = 15$

Where $e(A)$ and $e(B)$ represent the expected retweet count of A and B respectively.

In this case, both Tweets would have been flagged as 'interesting' under Proposition 4.1 (although, in reality, the derived method would not be able to model users who are typically expected to achieve 3,000 retweets). However, it is clear that, despite the *difference* between the counts being equal, Tweet B's observed retweet count is actually much more significantly proportionately greater than what was expected, and is therefore likely to be more significantly interesting.

Since the proportionate difference is the key to this, the interestingness score, $s(t)$, for

Tweet t is simply given by;

$$s(t) = \frac{t.\text{count}_R}{e(t)}$$

This provides a positive score where;

$$s(t) \begin{cases} > 1 & \text{indicates } t \text{ is interesting} \\ \leq 1 & \text{indicates } t \text{ is non-interesting} \end{cases}$$

and where $s(A) > s(B)$ implies that A is estimated to be more interesting than B.

Since this methodology relies on data collection from Twitter in order to obtain the observed retweet counts, it involves extracting a snapshot of the state of the evaluated Tweets at one stage during their lifetime. Since Tweets are not removed over time (unless they are deleted by their author) they can be discovered and retweeted at any time after their composition and posting.

The work in this chapter assumes that the most significant portion of retweet activity for a specific Tweet has already occurred by the time the information on the Tweet has been collected. Kwak et al. [35] carried out investigative analyses into various temporal retweet behaviours, and discovered that, on average, a Tweet receives around 75% of its retweets within the first day of being posted. 50% of the retweets of a Tweet even take place within the first *hour* of the Tweet being posted. Due to this, and to ensure that the retweet count collected is mostly representative of the Tweet's extrapolated 'final' retweet count, only Tweets that had been posted at least one day ago were considered for experimentation.

5.2 Further Adaptations of the Inference Methodology

Limitations with the previous method dictated that predicting a Tweet's expected retweet count could only work under certain restrictions. In particular, that the user must have a small enough local network (in practice, a follower count of more than 500 or so made the method very unsuitable), and that, due to this, Tweets only attracting very few retweets could effectively be simulated. In addition, Section 4.2.2 demonstrated that the validations of the interestingness inferences were found to be inaccurate in terms of the agreement with human judgments, although this is likely due to a combination of the above issue in providing much less room for error and the fact that the interestingness decision was only binary.

A new method is proposed, derived from Proposition 5.1, for carrying out the prediction for the value of $e(t)$. The method involves generating a classifier model capable of producing a base-line expected retweet count for a given Tweet and its relationship with its author. In this case, the classifier would be trained with the Tweet's actual retweet *count* instead of the binary retweet decision used previously, and it would not require the simulations of the user's local network. Many more features regarding the Tweet, and its content, and its author are used to represent the particular user-Tweet information required for generating the predictions.

Since the graph structure clearly has an impact on message propagation, then it was felt that a significant consideration should be made towards including features relating to the interconnection of users, such as follower counts, Tweet rate, and information on a sample of friends and followers. More detail on the features used is provided in later sections.

Experiments to demonstrate the newly-proposed methodology are to follow these steps:

1. Collect information on a number of Tweets and their respective authors from Twitter;
2. Form a dataset, T, of Tweets, each with its author's information embedded;
3. Split T into a training and test set - T_{train} and T^{a}_{test} respectively;
4. Train a classifier on $t \forall t \in T_{\text{train}}$. This trained model is known as the global model;

5. Extract features for each $t \in T^a_{\text{test}}$ and classify each against the trained classifier to obtain $e(t) \forall t \in T^a_{\text{test}}$;

6. Calculate $s(t)$ from the obtained $e(t)$ and the known $t.\text{count}_R$.

This method is immediately more superior as only a very small amount of data is required to be collected from Twitter. This means that inferences on Tweet interestingness could be made on demand[1].

In addition to the 'global' model, a 'user' model was proposed to be built for each user being evaluated. This user model would be much smaller, as it would be based only on the features from a history of that user's Tweets, but would be capable of providing a second value for $e(t)$ when testing Tweets against it.

Definition 5.1

A Tweet, t's, **expected 'global' retweet count**, $e_G(t)$, is the retweet count predicted when classifying t's Tweet and user features against the trained global model.
A Tweet, t's, **expected 'user' retweet count**, $e_U(t)$, is the retweet count predicted when classifying t's Tweet and user features against $t.\text{author}_O$'s trained user model.

With two such values generated through the comparisons of Tweets to both the global and user models, two scores could be generated as a function of the static value of $t.\text{count}_R$.

Definition 5.2

A Tweet, t's **global score** is derived from the global model and is denoted as $s_G(t) = \frac{t.\text{count}_R}{e_G(t)}$.
A Tweet, t's **user score** is derived from $t.\text{author}_O$'s user model and is denoted as $s_U(t) = \frac{t.\text{count}_R}{e_U(t)}$.

[1]Not 'live' as it relies on time for retweets to have occurred.

5.3 Collecting the Training and Testing Data

In order to train the model on a set of Tweets and then use it to make predictions, data was required for collection from Twitter. This data is relevant to the experiments and analyses in the following sections.

Since, in this case, it was necessary to collect the Tweet data along with each Tweet's numeric retweet count, rather than the binary nominal yes/no required in the previous chapter, only the retweets of a particular Tweet that had been created using the button method could be considered. This is because a Tweet's retweets executed using the manual copy and paste method do not contribute to the Tweet's official, and observable, retweet count that is returned from Twitter's API. This is not considered to be a limitation, however, since this factor is considered consistently through the training and later evaluation of the trained model.

In March 2013, a random walk was conducted through Twitter's social graph using v1.1 of Twitter's REST API. Although this date was before the mandatory transfer to this version of the API, the crawler method was used in preference to collecting from the public timeline, which was deprecated and removed in v1.1, so that user data could be collected to account for the importance of the social graph in information propagation.

Each stage of the walk consisted of focusing on and collecting information on a user, u. As such, the crawler is very similar to that used in the latter sections of the previous chapter. At each stage of the crawl, a set of recent Tweets, T_u, where $t.\text{author}_O = u \forall t \in T_u$ were collected. The size of T_u had various dependencies, such as the u's Tweet-posting frequency and the number of Tweets in total authored by u. Usually, several hundred Tweets from each user were yielded. In addition to the Tweet data, information on u itself was collected as well as on a sample subset of up to 100, if they exist, of each of $N^+(u)$ and $N^-(u)$. The next stage involved focussing on a new user selected randomly from $N^+(u)$. Where 'dead-ends' occurred (in cases where

$\deg^+(u) = 0$ or the case where all $N^+(u)$ having already been collected), the crawler back-tracked to the most recently-collected user from which to select a valid follower to continue the crawl.

The sample subset of friends and followers of each user was collected instead of the complete set for the purposes of efficiency and to address the associated limitation in the previous interestingness inference methodology. However, the samples still provide an example snapshot of up to an additional 200 users in the author's neighbourhood in order to provide some idea of the activity within the local network both upstream and downstream from the author user. Around ten API calls were required to obtain this information for each user, giving it immediate advantages over the older method, which required several hundred or thousand depending on the particular user.

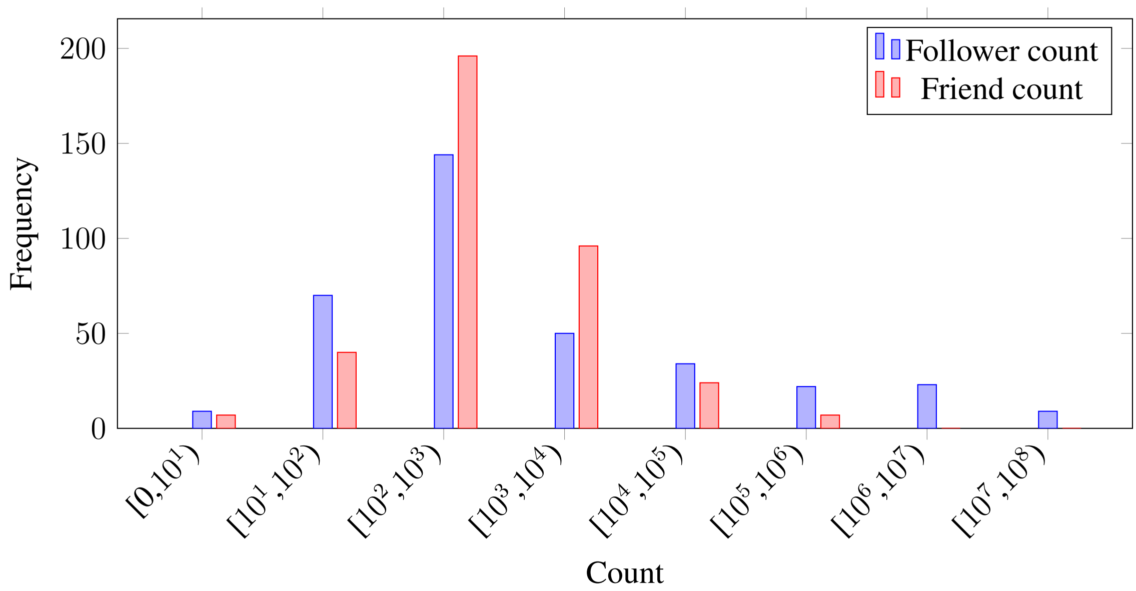


Figure 5.1: Distribution of follower and friend counts of authors of Tweets in dataset T.

Importantly, Tweets from many different types of user were collected; from less-active users with very few followers and friends to influential users and celebrities with millions of followers and achieving many thousand retweets. The distribution of the follower and friend counts of the authors of Tweets in T is displayed in Figure 5.1 and the collection of this wide range of users will help demonstrate if the new methodology is able to assess a larger variety of users and Tweets.

The data collection resulted in a set of Tweets, $T = T_u^1 \cup T_u^2 \cup ... \cup T_u^n$, where $|T| = 240,717$ and where $n = 370$. Of T, around 90,000 were Tweets with a retweet count of greater than zero. While the data was being collected, parts of the complete set were used in analyses of classifiers and Tweet categorisation methods to be used in the forthcoming experiments. $T_{p1} \subset T$, where $|T_{p1}| \approx 57,000$, represents the Tweets in T part way through the data collection and is used in Section 5.5.2, and $T_{p2} \subset T$, where $|T_{p2}| \approx 67,000$, is used for the analyses in Section 5.5.3. The entire set is used in the main validation analyses later on.

5.4 Retweet Counts as Nominal Attributes

Many machine learning classifiers are not able to accurately predict the outcome of a feature of a large-ranging and contiguous data type. Since retweet counts are on a very wide-ranging contiguous scale, using a classifier to make predictions from a limited range of discrete ranges, or 'nominal' data, would be more appropriate.

Thus, in order to help improve the accuracy of $e(t)$ predictions, it was decided to convert the retweet count feature into a nominal data type for the purposes of training the model and making classifications. By 'binning' the retweet counts into categories representing interval ranges, there would be fewer outcome possibilities, and thus the *confidence* of classification could be greater.

The values for $s(t)$ would then be determined through the ratio of $t.\text{count}_R$ to the *upper* limit of the category containing $e(t)$.

Generally, trained classifiers are only able to make predictions on features and values it has prior knowledge of. Therefore, the bin ranges for each category must be equal in both the training feature data and the testing feature data. If the available nominal values for an instance feature representing a Tweet has a different set of category ranges to that in the trained classifier model, then it is likely that a prediction cannot be generated for this instance. This is an important factor to consider when determining a method for binning the retweet counts.

One method to create bins based on a set of contiguous values is to do so in a *linear* fashion. Given a set of retweet counts, A, where $A = \{t_1.\text{count}_R, ..., t_n.\text{count}_R\}$, then the range of retweet counts, $c_r = \max(A) - \min(A)$. The linear approach then determines the range-size of each bin to be $\frac{c_r}{B}$, where B is the desired number of bins. A particular retweet count, $t_i.\text{count}_R$, can be assigned to a bin, b, if $x \leq t_i.\text{count}_R < y$, where the interval describing $b = [x, y)$. The list of bins produced by the methodology represents the available nominal categories that each Tweet's retweet count can be assigned to.

In cases where $\min(A) \neq 0$, the interval $[0, \min(A))$ is pre-pended to the list of bins. Similarly, in all cases, the interval $[\max(A) + 1, \infty)$ is appended to the list. This dictates that no Tweet in the set from which A is derived can have a value for $t.\text{count}_R$ categorised into this bin, and thus this allows any Tweet to potentially have $s(t) > 1$ when testing against the eventual model. For example, if a training set of Tweets with a total range of values for $t.\text{count}_R$ being between 1 and 20 was binned into four ranges, then the following interval categories would be applicable:

$$[0, 1), [1, 6), [6, 11), [11, 16), [16, 21), [21, \infty)$$

Since the distribution of retweet counts (expressed through retweet group sizes) is known [57], then it is clear that this binning methodology would produce bins containing a very non-uniform distribution of Tweets, where the lower bin ranges would contain many Tweets and the cardinality of each category would decrease exponentially as the categories become higher. Figure 5.2 illustrates the non-uniformity of the resultant bin sizes for the set of Tweets, T, linearly binned even with an exaggerated $B = 30$. The latter intervals are combined in the Figure for readability. The result of

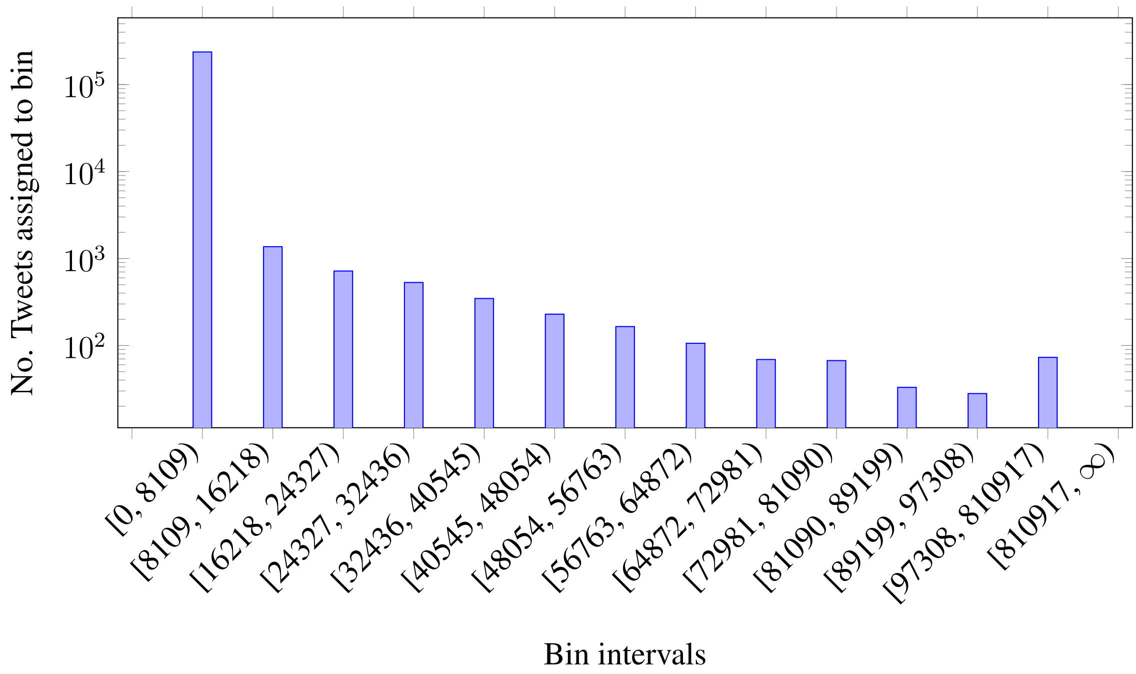


Figure 5.2: Bin intervals and cardinalities for the retweet counts of Tweets in T linearly binned with $B = 30$.

this means that there are significantly fewer feature instances representing Tweets with larger retweet counts.

Indeed, when training a Bayesian Network classifier and using it to run cross-validations on the retweet counts of Tweets in T, this binning scheme demonstrated a high accuracy of predictions on Tweets with lower values for $t.\text{count}_R$ and a low accuracy for Tweets with higher counts. Table 5.1 shows this high precision and recall for the bin containing the majority of the Tweets, along with the very poor performance of the subsequent bins. Only the first ten bins are shown, since the remaining 20 bins produced precision and recall values of 0. It would be more appropriate, and better address the desire for more universal use-cases expressed earlier in this and the previous chapter, if the accuracy of predictions could be more uniform across the bin ranges.

A responsive approach dependent on the range and distribution of retweet counts would help in producing more evenly-filled bins and therefore increase the prediction per-

Retweet count bin interval	Precision	Recall
[0, 8109)	1.0	0.956
[8109, 16218)	0.083	0.355
[16218, 24327)	0.134	0.315
[24327, 32436)	0.233	0.072
[32436, 40545)	0.0	0.0
[40545, 48654)	0.008	0.004
[48654, 56763)	0.105	0.109
[56763, 64872)	0.03	0.038
[64872, 72981)	0.008	0.174
[72981, 81090)	0.009	0.343

Table 5.1: Cross-validation performance results for the first 10 bins produced through the linear binning method on retweet counts in T with $B = 30$.

formance across the range of intervals. A new methodology is proposed, following a retweet count distribution based approach, which uses the size of the Tweet set to be categorised and the bin count B for dynamically generating interval ranges such that the cardinalities of each bin are as even as possible. Each bin can then be filled according to the bounds of its interval, and in such a way as to ensure that each retweet count frequency would only be present in one bin. For example, all of the retweets achieving one retweet would be placed in the single bin encompassing this value. Algorithm 2 illustrates this dynamic approach in more detail.

As such, after the intervals covering the bin bounds have been produced, then these represent the nominal categories for the retweet count feature in each instance for training and testing against the classifier.

Algorithm 2 Algorithm for producing intervals for bin categories for $t.\text{count}_R$ values.

procedure GENERATE_INTERVALS(set of Tweets T, number of bins B)
    $C \leftarrow$ empty list ▷ To hold ordered retweet counts
    $I \leftarrow$ empty list ▷ To represent bin range intervals
    for all $t \in T$ **do**
        Add $t.\text{count}_R$ to C
    end for
    Sort C into ascending order
    $M \leftarrow \max(C)$ ▷ Highest instance of $t.\text{count}_R$
    $T\text{Sum} \leftarrow \lceil \frac{|C|}{B} \rceil$ ▷ Number of Tweets to be held by each bin
    $H \leftarrow$ empty dictionary ▷ To represent the distribution of retweet counts

    for all $c \in C$ **do**
        if $c \in H$ **then**
            Increment H_c
        else
            $H_c \leftarrow 1$
        end if
    end for
    for all i in range $1, ..., M+1$ **do**
        if $i \in H$ **then**
            $s \leftarrow s + H_i$
        end if
        if $s \geq T\text{Sum}$ **then**
            Add i to I
        end if
    end for
    Return I
end procedure

This method readily supports more uniform bin sizes, and copes with this by exhibiting exponentially larger bin *ranges*. As with the linear method, the interval $[0, \min(A))$ is pre-pended, where the dataset requires it, and $[\max(A) + 1, \infty)$ is always appended in addition to the intervals produced by the algorithm. The distribution of bins for the retweet counts for the Tweets in T when binned through this method is illustrated by Figure 5.3. The same logarithmic scale is used as in Figure 5.3 to allow the comparison to help illustrate the greater uniformity in this dynamic method.

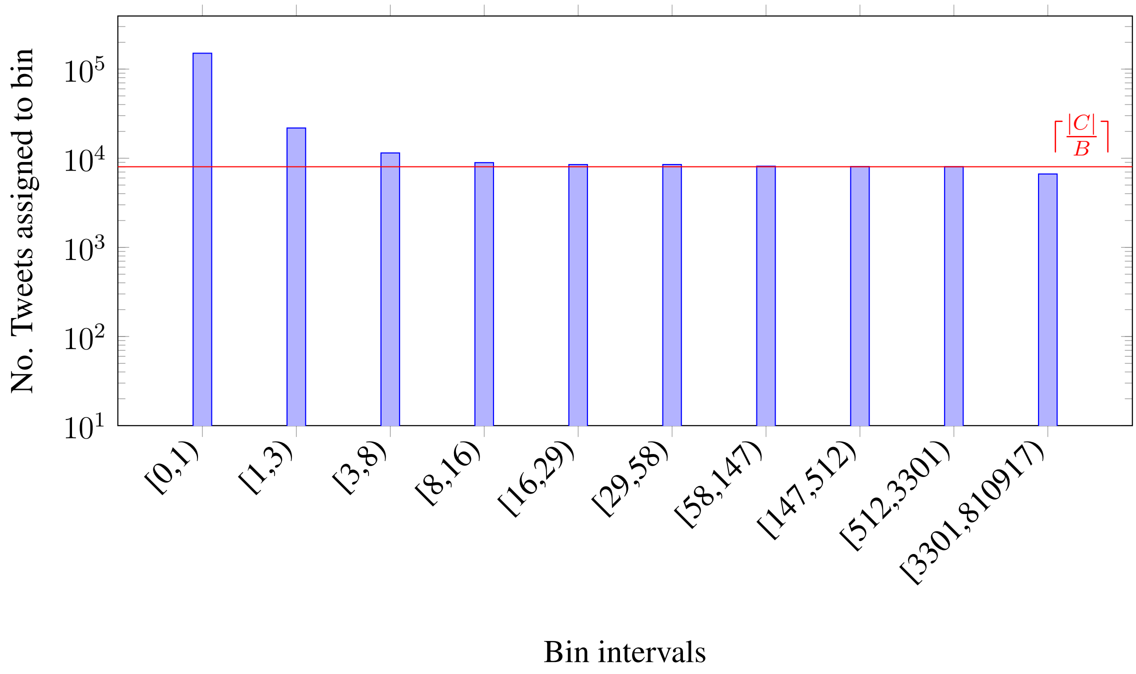


Figure 5.3: Bin intervals and cardinalities for the retweet counts of Tweets in T dynamically binned with $B = 15$.

The responsiveness stems from the fact that the bin ranges adapt to the variety and number of retweet counts present, and the method always attempts to produce a similar number of bins to the target count. However, due to the disproportionately large number of small retweet groups, the bin sizes cannot be entirely uniform and this means that the number of bins returned will generally be smaller than the target number. Furthermore, a single Tweet cannot exist in more than one bin concurrently. In the case of Figure 5.3, the number of Tweets with retweet count in the interval $[0, 1)$ is greater than $\lceil \frac{|C|}{B} \rceil$, where C is the ordered list of retweet counts of Tweets in T, resulting in a significantly larger first bin and an overall bin count of less than B. Without this particular feature, training Tweets with equal observed retweet count may be categorised into multiple bins, which may cause complications with the training and eventual testing of the model.

The performance given by cross validations of a Bayesian Network classifier on the retweet counts of Tweets in T is given by Table 5.2. It is clear that the precision and recall accuracy is much more evenly balanced across the bin intervals generated by the dynamic method than those by the linear method, the latter of which were reported in Table 5.1.

Retweet count bin interval	Precision	Recall
[0, 1)	0.935	0.741
[1,3)	0.218	0.324
[3,8)	0.190	0.394
[8,16)	0.240	0.233
[16,29)	0.291	0.298
[29,58)	0.265	0.338
[58,147)	0.232	0.201
[147,512)	0.256	0.418
[512,3301)	0.527	0.508
[3301,810917)	0.519	0.709

Table 5.2: Cross-validation performance results for the first 10 bins produced through the dynamic binning method on retweet counts in T with $B = 15$.

Due to this dynamicity, the bin ranges and cardinalities produced by the algorithm vary across different datasets. As a result, the nominal bin categories generated for producing the value for $e_G(t)$ from the user model trained from the complete set of collected Tweets posted by $t.\text{author}_O$ would be different from those categories generated for a different user. The intervals in each bin category are therefore reflective of the various numbers of retweets that each author's Tweets are likely to receive.

Thus, when testing against either the global or user models, the Tweets' retweet counts are predicted as *nominal* counts equal to the upper bound of the bin category the Tweets are classified as. For example, consider a Tweet, t, to be tested against a classifier that was trained by a training set of Tweets that produced the bin intervals shown in Table 5.2. This Tweet has $t.\text{count}_R = 20$, but is, in this example, classified as $e(t) = [8, 16)$, and therefore receives a score of $s(t) = \frac{20}{16} = 1.25$.

Similarly, when training, the bin intervals are first generated from the training set's retweet counts. Each Tweet then has its retweet count transformed into the generated nominal bin interval that surrounds the Tweet's retweet count before the model is trained. If the same training set is used as in the previous example and one of the Tweets in the set has a retweet count of 48, then this Tweet has its count transformed into the nominal [29,58) for training with.

5.5 Predicting Estimated Retweet Counts

Since the environment has previously been found to have a large effect on propagation, then the features describing a Tweet's author's relationships and local network activity are a useful aid in feature selection, as described by references to memetics in the Background chapter.

In order to generate the estimated retweet counts, a trained machine learning classifier is used to make predictions on a set of instances made up from both environmental and Tweet features. This section covers the process of selecting a classifier and justifying its use in terms of analysing its performance.

5.5.1 The Classifier

An overview of machine learning classifiers and their processes was provided in the previous chapter. In that case, a logistic regression was used to generate a prediction on a binary retweet decision based on a small number of features. If the retweet count for the Tweet being trained or tested was greater than zero, then the retweet decision would be positive (TRUE). Otherwise, the decision was negative (FALSE).

The improved methodology proposed in this chapter involves the prediction of a retweet count category from a set of nominal values of cardinality greater than one. As mentioned, the instances of a particular Tweet and its environment are categorised based on the value of the retweet count of the Tweet. Although this means that a degree of accuracy is sacrificed when training the classifier, it does mean that there are fewer categories for predictions on test Tweet feature instances, providing a higher confidence for each prediction made.

A Bayesian network machine learning classifier was elected for use for the purposes required in this chapter. Use of this classifier in the social media domain is more rare than other classifiers, such as those involving a regression or a decision tree, but was selected due to its performance and efficiency shown in Table 5.3.

The Bayesian network is an unsupervised classifier since its learning algorithms do not simply determine the class of the outcome, the retweet count, from the attribute features alone [21]. Instead, a probabilistic graph is constructed based on the dependencies between the variables. The variable attributes form the nodes of the graph and edges between the nodes denote the dependencies (or lack thereof) between them.

In the case of this research, the various Tweet and environmental features, including the nominal retweet count, form the nodes in the Bayesian Network. When forming the network through training, the dependencies and their probabilistic weightings are adjusted so that an expected value for the retweet count can then be 'predicted' from the values of all the other variable attributes.

5.5.2 Classification Performance

The choice of classifier stems from its combined efficiency and accuracy when working on the data relevant to this chapter. This section provides a brief overview and comparison of a subset of commonly-used (and related) classifiers in social network research.

Choosing a classifier

In order to study and evaluate the relative performance of appropriate classifiers for selection for this task, the Weka[2] machine learning and data mining toolkit was used. The classifiers were selected to cover a sample of the range of available classifier categories. Whilst some types may work inefficiently in this scenario, it is likely that they are more efficient when employed in different use-cases.

Although the accuracy of prediction was important, it would also be useful for the classifier to be *efficient* in training its model and when testing future instances against it. This is so that this method could be used to produce interestingness inferences with a more on-demand policy and to further improve on the methodologies used in the previous chapter.

Classifier	Weighted average precision	Weighted average recall	Training time (secs.)
Simple logistic	52%	56%	528
Logistic	62%	56%	18
SMO	51%	55%	1384
Naïve Bayesian	50%	44%	0.13
Bayesian network	62%	64%	0.54

Table 5.3: The performance of different machine learning classifiers in cross-validations on dataset T_{p1}.

[2]http://www.cs.waikato.ac.nz/ml/weka

Table 5.3 shows the Bayesian network to be accurate and time efficient when evaluating the performance of the set of evaluated classifiers when running cross-validations on the dataset T_{p1}, which is a subset of T and contains around 57,000 Tweets. For each Tweet instance the same retweet count binning scheme was used, and each classifier performed the same number of cross-validations against the same dataset in order to obtain the precision and recall values.

Although the dataset used in this analysis is not the complete set used in practice, the cardinality of the dataset was sufficient to cause the outputs to be indicative of the Bayesian network's relative advantages over the other assessed classifiers.

Overview & evaluation of classifier performance

The data observed in Table 5.3 illustrate how different algorithms approach classification in various ways. In addition to the *type* of the data in the instance features, the cardinality of the dataset can also impact training efficiency. Following is an overview of the different classifiers assessed. 'Independent variables' represent the input values to be classified in each case, and a 'dependent variable' is the output class given by the inputs.

Simple logistic

Linear regression involves the prediction of a dependent scalar value derived from a set of independent variables. For example, consider the case of one dependent, one independent variable, and their modelled relationship. Previously-seen values for the independent variable can be used to estimate an associated value for the dependent variable. Predictions for the dependent variable with unseen independent instances can be inferred from surrounding 'known' instances.

Simple logistic regression is analogous to linear regression, except the dependent variable is nominal, and the regression allows for producing probabilities that the independent variables belong to a particular class [53].

Logistic

As with its 'simple' counterpart, logistic regression involves the prediction of a nominal variable derived from a set of independent values. Analyses in this scope are categorised as binary (two nominal classes) or multinomial (three or more binary output classes), and the output probability is produced through the natural logarithm of the dependent variable being the case given the independent variables.

SMO

Sequential minimal optimisation (SMO) is an algorithm for producing support vector machines (SVMs) quickly and cheaply [48]. SVMs are non-probabilistic graph models for nominal classification, and work by representing independent variables in space, but separated into clusters representing the class of the dependent variable. The output class of new independent variables is given by the cluster they are mapped to.

Naïve Bayesian

Naïve Bayesian classification is able to consider many multi-dimensional (including nominal and scalar) features. In particular, Bayesian probability is applied towards prediction of the value of the dependent variable, yet *without* considering any relationships or weightings between the independent variables. For example, a water bottle is long, thin, and cylindrical. A naïve Bayesian classifier trained on these independent variables may still classify instances as being a water bottle if there is enough evidence to suggest it is cylindrical, even if it is *not* long and thin.

Bayesian network

A Bayesian network classifier involves a graph model, whose nodes represent the independent and dependent variables and whose edges indicate the influences between variables. Upon training and testing, a Bayesian probability distribution is generated over the variables, which is used to indicate the resultant class of the dependent variable given the relationships between the independent variables and itself.

Evaluation

As seen in Table 5.3, the two Bayesian probability-based approaches are significantly faster than the logistic and support vector machine methods. Classification through logistic regression involves further searches through its matrix model upon addition of training instances[3], which takes longer to complete as training data size increases. Similarly, when producing the support vector machines, the sequential minimal optimisation algorithm must normalise the input features as new instances are added.

The Bayesian network is able to model relationships between indepdendent variables, and the naïve Bayesian classifier is not. This means that these relationships allow for representation of the ties between features in the underlying social graph, such as observed relationships between follower and friend count for 'measuring' influence. Therefore, they can be used to more accurately infer a projected retweet count than when they are not considered, and the more complex computation accounts for the quicker training time achievable with the naïve approach.

5.5.3 Effects of varying the Cardinality of Nominal Retweet Counts

Applying the continuous retweet count values to produce a set of nominal categories representing interval ranges of the retweet counts requires a certain balance. By reducing the number of target category bins then the classification accuracy increases, but the level of applicability of the eventual interestingness score for the wide range of retweet counts observed would be reduced since the *granularity* of the predicted counts would also decrease. Conversely, with more bins, the classification accuracy reduces, as there would be fewer instances in each category, yet the scores would be apliccable to a wider range of retweet counts.

Clearly, by increasing the number of nominal categories used, then the relative number of feature instances in each eventual interval decreases. These bins represent the nominal categories that each feature instance is classified as in relation to the predicted retweet count of the instance. Table 5.4 outlines the decrease in classification accuracy

[3]http://www.academia.edu/5167325/Weka_Classifiers_Summary

Target bin count (B)	Resultant bin count	Weighted average precision	Weighted average recall
1	1	100%	100%
2	2	89.3%	89.3%
5	4	78.8%	74.5%
10	7	68.6%	65.7%
15	10	61.2%	56.4%
20	12	59.1%	52.9%
25	15	51.4%	47.5%
30	18	49.3%	45.3%
35	21	47.2%	43.2%
40	23	46.2%	42.5%

Table 5.4: The effect of varying B on the cross-validation performance using a Bayesian network classifier on dataset T_{p2}.

observed with increases in target bin count using the dataset T_{p2}.

In the upcoming experiments a value of $B = 14$ was used, which yielded ten nominal retweet count categories for use in training and testing against the general global dataset for the purposes of generating the global expected retweet count. Since each user's own retweet count ranges were different, the number of categories were calculated individually for producing the user-centric expected retweet counts as part of calculating values for $s_U(t)$.

5.6 Training and Testing Against the Classifier

This section discusses the processes used to calculate interestingness scores for Tweets through the generation of expected retweet counts using the methodologies outlined in the previous sections. Particular focus is lended to the managing of the data corpora and feature extraction.

5.6.1 Data Corpora

The dataset T was collected, as described earlier, and is used for training the model and for validity testing. As such, T was divided into two datasets; a training set, denoted T_{train} and consisting of 90% of the entire set, and a testing dataset, denoted by T^a_{test} and consisting of the remaining 10%. T was divided in such a way as to ensure that all of the Tweets authored by one particular user existed in only one of the two resultant datasets. After being used to train the Bayesian network model, T_{train} was then discarded from use for the rest of the experimentation.

In order to support the user scores, in addition to the global scores, further datasets are required to be extracted from T^a_{test} to produce an individual T^u_{test} for each $t \in T^a_{\text{test}}$ where $t.\text{author}_O = u$. Each of these subsets contains the Tweets written by and information about that particular user, and can be used to train an individual Bayesian network classifier specific to the user. T^a_{test} is therefore referred to as the 'global' Tweet corpus for testing against the global model to generate global Tweet scores, and T^u_{test} is referred to as the 'user' Tweet corpus for user u in order to train that author's user model and for testing the Tweets it posts to generate *user* Tweet scores. Each Tweet, t, therefore, can be evaluated against both the global model and its author's user model in order to respectively produce values for $e_G(t)$ and $e_U(t)$.

It should be noted that the same user corpus is used for both training the model and for testing against in the validations of the user scores. This is due to the relatively low number of Tweets available for training each individual user model, and so re-use of the Tweets is necessary for the research in this chapter. The process for generating the global scores still maintains distinct datasets for training (T_{train}) and testing (T^a_{test}).

5.6.2 Features

Producing the instances used for testing and training the Bayesian network models involved the extraction of various features from the global and user datasets. Generally, each feature falls into one of three categories; the network features ('environment'), the Tweet features ('genome'), and the author features (representing the author of the current Tweet). The nominalised retweet count is categorised as a Tweet feature.

Generally, the Tweet features follow the same notions as those used in the previous chapter in that they are static and generally binary features describing various aspects of the Tweet's content and metadata. The network features are more variable and describe the ways in which the author's local network is constructed and the activity within it.

Each Tweet is represented by an instance of a complete set of features relating to that Tweet, its author, and its author's local network. As a result, feature instances representing Tweets authored by the same user will share the same values for their network and author features.

Features for the global corpus model

The global corpus model is the Bayesian network model representing the classifier trained from the complete training dataset. In this case, a total of 31 features, outlined in Table 5.5, were used to train the classifier. As such, there were around 217,000 Tweet instances using this feature scheme used for training the global classifier.

The network features listed apply to both samples of the followers and friends retrieved for each author user during the data collection. For example, the first feature of this category, 'max. follower count', represents two features referring to the maximum follower count observed across the sample of the user's followers and the sample of the user's friends respectively.

It should be noted that although the Tweet features, aside from the retweet count as has already been discussed, are permanent after the Tweet has been created and posted, the author and network features are more dynamic due to the continuous mutations in the social graph as edges representing followships are constantly being formed and

Feature category	Feature	Feature data type
Tweet ('genome')	mention	{True, False}
	Tweet length	real (numeric)
	url	{True, False}
	hashtag	{True, False}
	positive emoticon	{True, False}
	negative emoticon	{True, False}
	exclamation mark	{True, False}
	question mark	{True, False}
	starts with 'RT'	{True, False}
	is an @-reply	{True, False}
	retweet count	**[dynamic nominal]**
Author	follower count	real (numeric)
	friend count	real (numeric)
	verified account	{True, False}
	status count	real (numeric)
	listed count	real (numeric)
Network ('environment')	max. follower count	real (numeric)
	min. follower count	real (numeric)
	avg. follower count	real (numeric)
	max. friend count	real (numeric)
	min. friend count	real (numeric)
	avg. friend count	real (numeric)
	avg. status count	real (numeric)
	proportion verified	real (numeric)

Table 5.5: Features used to train the model from the global data corpus.

broken between the user nodes. In this thesis, it is assumed that changes to the features representing these factors were not significant over the period of posted Tweets for each user, and the effect is minimised through consideration only of the recent Tweets of each author user.

After training the classifier with these features from the set T_{train}, each Tweet, $t \in T^a_{\text{test}}$

was tested against the model in order to classify it into a retweet outcome category, as described above. The upper bound of this category interval is then used, along with $t.\text{count}_R$ to assign t a numeric global score, $s_G(t)$.

Features for individual user models

Since the author and network features have identical values in the instances representing all of the Tweets from one particular user, then these features were not considered when training and testing using the user models. As such, the 10 Tweet features were those used in the feature instances in training, and testing against, each user model.

After training each user classifier with the features representing that particular user and its Tweets, each Tweet $t \in T^a_{\text{test}}$ was tested against the classifier model representing the features of $t.\text{author}_O$ in order to assign it a numeric user score, $s_U(t)$.

5.7 Initial Validations of the Scoring Methodologies

In order to verify the accuracy of the assignment of both of the scores to each Tweet in T^a_{test}, validation tests by human participants was required. Through running these validations, the relative performance of the scoring mechanism can be assessed, and the comparative performance of the two scores, $s_U(t)$ and $s_G(t)$ can be evaluated.

Mechanical Turk was again used to crowdsource inputs from MTWs, as this would facilitate the obtaining of interestingness evaluations from a wider range of human opinion.

5.7.1 Planning the Validations

The MTWs taking part would not be associated with the collected Tweets in any way, and thus this assists in the identification of the non-noisy Tweets that are ‘globally’ interesting and are those that the scores have theoretically determined as 'interesting'.

At this stage, certain Tweets and users were removed from T^a_{test}. Since the Tweet and user data was collected as part of a random crawl, there was no governance over the content of the Tweets collected. As such, users who frequently used offensive phrases or did not write Tweets in English had their Tweets removed from T^a_{test} for the validations, since the Mechanical Turk microtasks were submitted to be assessed by MTWs from the USA. As before, individual Tweets that were '@-replies' were also removed so that only Tweets intended for broadcast were included in the final set to be evaluated.

5.7.2 Carrying Out the Validations

In the context of this validation scheme, MTWs were not required to hold an active Twitter account. By not determining the humans to make the assessments, a more diverse opinion on the interestingness can be achieved, as the different users will have varying considerations on what constitutes 'noise' and will therefore reinforce a decision when multiple MTWs form agreements on what is interesting.

The validations were carried out such that the MTWs were presented with a series of questions, each of which consisting of five different Tweets from *one* specific author. An example question is shown in Figure 5.4 and, as such, Tweets were assessed against others that had been posted by the same user. In each question, the MTWs were asked to select the Tweets that they consider to be the most interesting of the group, and that they must select at least one Tweet for each question. For each judgment, where a judgment is one question answered with one or more Tweets selected, MTWs were paid \$0.05. Given the relative shortness of each Tweet, each judgment was expected to take less than 30 seconds to complete (including quickly visiting URLs in those Tweets that contained them), meaning that MTWs would be paid an average of at least \$6.00 per hour if completed continuously. This rate is generally acceptable for workers

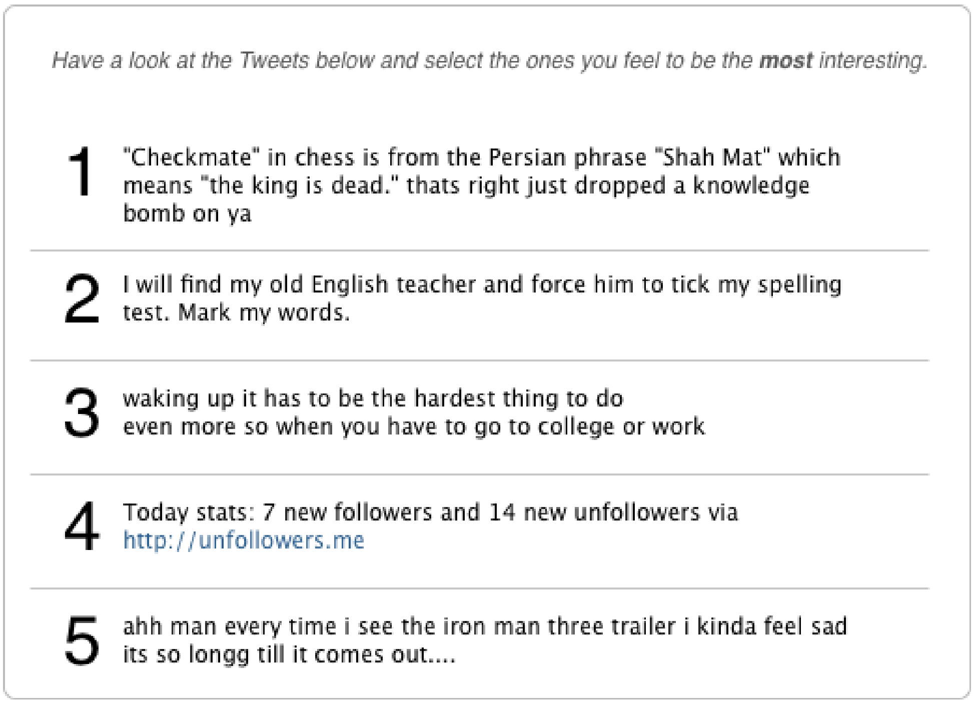


Figure 5.4: An example validation question.

completing this type of easier, multiple-choice task and for newer workers[4].

The test was conducted under the conditions of a randomised controlled trial. To this end, each Tweet was assessed in three different contexts, in that it would appear in three different questions alongside four other randomly chosen Tweets, and that each question would then be judged by three different MTWs.

From the stripped testing dataset, 750 Tweets were selected, at random, to be divided by user into the questions to be assessed on Mechanical Turk. Since each Tweet was to appear in three different questions and since each question consisted of five unique Tweets, then this resulted in a total of 450 distinct questions. Each Tweet was assessed as part of a question nine times in total.

[4]http://blog.echen.me/2012/04/25/making-the-most-of-mechanical-turk-tips-and-best-practices

5.7.3 Outcomes From the Validations

The validation test involved contributions from 91 different MTWs in total, and 325 of the 450 questions in total asked had responses where a Tweet was selected *confidently*. A confidently-answered question is defined as the case when at least two of the three respondent MTWs answering that question selected the same Tweet. Since the MTWs had the opportunity to select more than one Tweet of each question to be the most interesting, there were 349 Tweets of the original 750 Tweets, denoted as $T' : T' \subset T^a_{\text{test}}$, that were selected as sufficiently interesting by the MTWs. Tweets selected from individual questions that did not have sufficient confidence were discarded.

The remainder of this section analyses the validation data in various ways to demonstrate the strengths and weaknesses of the interestingness score inferences. Of immediate notice was the comparative difference between the two different scoring mechanisms for each Tweet t; $s_G(t)$ and $s_U(t)$. The inference validation results are not significant between the use of the two scores in any of the analyses conducted.

General Performance

Of the subset T', the scoring mechanism found 140 of the Tweets to have a value of $s_G(t) > 1$, and thus inferred as *more* interesting than the remaining Tweets. Of these, 65% were agreed on as interesting by the MTWs. The performance of the $s_U(t)$ was worse in providing a 55% agreement, resulting in a general of 60% agreement on the mean of the two scoring schemes.

Observation 5.1

Although not significant, the general performance accuracy is demonstrably greater when using the **global** scoring scheme than the **user** scores.

It is also the case that the proportionate frequency of Tweets with higher values of $s_G(t)$ is greater in the subset T' than in T. This implies that, on average, Tweets selected by MTWs have a higher score than those not selected, as is shown in Figure

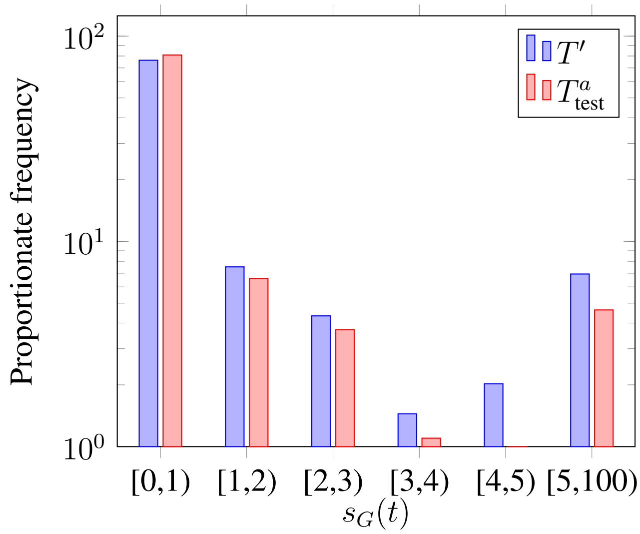


Figure 5.5: Proportionate frequency distribution of $s_G(t) \forall t \in T^a_{\text{test}}$ compared to only those $s_G(t) \forall t \in T'$.

5.5. Additionally, a greater proportion of Tweets with $s_G(t) < 1$ were in T^a_{test} than in T' and a greater proportion of Tweets with $s_G(t) > 1$ were in T' than in T^a_{test}.

Observation 5.2

MTWs select, on average, more Tweets with an 'interesting score' ($s_G(t) > 1$) than Tweets with a 'non-interesting score' ($s_G(t) < 1$).

Ranking Performance

In order to assess the ability of the scores to effectively rank Tweets in order of inferred interest *level*, the Tweets were studied on a per-question basis.

Let q, which represents a particular question containing five Tweets, $t_i^q \quad \forall \quad 1 \leq i \leq 5$, be ranked in order of ascending interestingness score such that;

$$q = (t_1^q, t_2^q, t_3^q, t_4^q, t_5^q)$$

where;

$$s_G(t_1^q) \geq s_G(t_2^q) \geq ... \geq s_G(t_5^q)$$

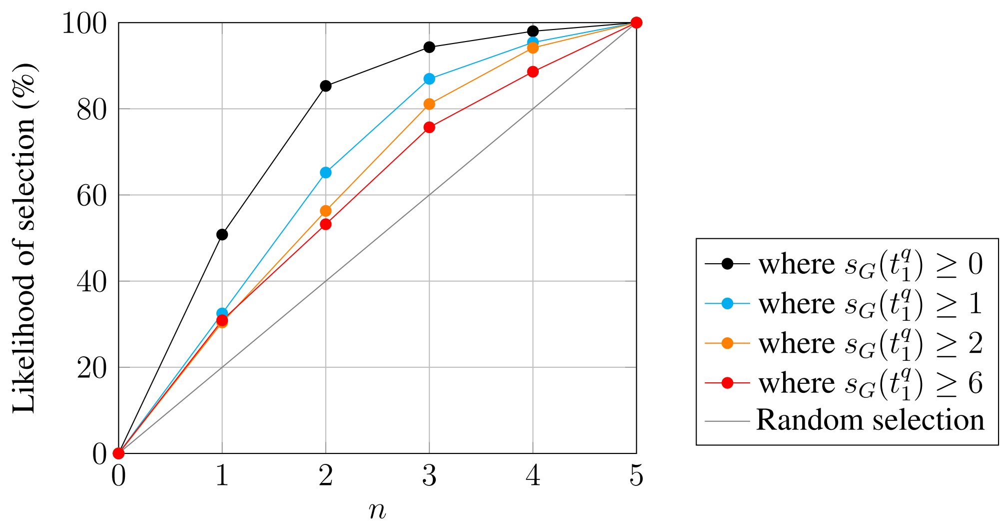


Figure 5.6: Likelihood of MTWs selecting one of the top n Tweets ranked by $s_G(t)$ in each $q \in Q$. Also illustrating the effect of raising the minimum allowed $s_G(t_1^q)$.

Let Q be the set of 450 such ordered questions asked of the MTWs, such that;

$$Q = \{q_1, q_2, ..., q_{450}\}$$

In cases where $\sum_{j=1}^{5} s_G(t_j^{q_i}) = 0$, q_i is removed from Q as none of $t_1^{q_i}, ..., t_5^{q_i}$ have been ranked as more interesting than one another or as interesting at all.

Given that the Tweets in each question, $q \in Q$, have been ranked in order of inferred interestingness, of relevance is the likelihood of MTWs selecting one of the first n of the Tweets in q. For example, consider the case of $n = 2$ with a set of ten questions, Q'. If the MTWs selected Tweet t_1^q or t_2^q in five instances of $q \in Q'$, then the average likelihood of selecting one of the top $n = 2$ Tweets for all $q \in Q'$ is 0.5.

Figure 5.6 illustrates the relationship between increasing values of n and this calculation on likelihood of selection. Although the 'random' performance represents the relative likelihoods of a random selection being made when only one Tweet is selected from each question, the vast majority of questions were actually answered with only one Tweet selected. Further analysis to cover the consideration of this particular point is conducted later on in this chapter.

When considering cases where the most interesting Tweet in a particular question is, indeed, inferred as interesting ($s_G(t) \geq 1$), then the MTWs selected one of the *top two* Tweets in around 66% of cases, and they selected one of the *top three* ranked Tweets in 87% of the questions. This demonstrates that the method's ranking ability is in agreement with humans in identifying information from the noise around them.

Observation 5.3

> MTWs select one of the top two highest-scoring Tweets in 66% of the questions in Q that contain at least one interesting Tweet ($s_G(t_1^q) \geq 1$).

Does Probability of Selection Increase with Tweet Score?

The relationship between the likelihood of selection and a Tweet's score is also of interest. It was found that, in general, the chance of a particular MTW deciding that a Tweet, t, is interesting becomes greater as the value of $s_G(t)$ increases.

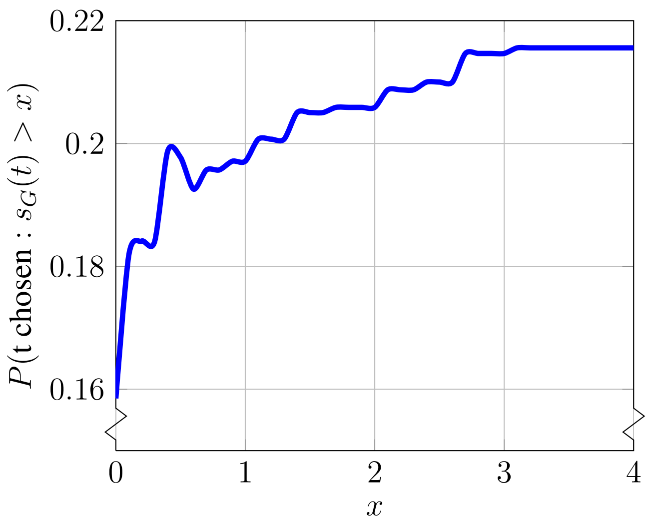


Figure 5.7: Cumulative frequency representing the probability that Tweet t is chosen provided that $s_G(t)$ is greater than a given value, x. Note that probabilities for $s_G(t) > 4$ have been excluded due to fewer samples.

Cases of Tweets where $s_G(t) > 4$ are excluded from this analysis, for the purposes of noise reduction from fewer samples, however Figure 5.7 shows an observable increase in probability of selection as the score increases. This pattern is particularly evident in the score interval of [0,1], which represents the range of Tweets that the scoring scheme has inferred as uninteresting to those that achieved a correctly-predicted popularity, and are thus 'as expected' in terms of interestingness. The analysis is also clear that Tweets with an inferred interestingness score of three or more are not significantly different from one another in terms of the level of interestingness assigned from the 'real' human judgment.

Is Interestingness More Identifiable with Greater Score Separation?

The metrics behind the human selection in determining interesting Tweets is the final analysis conducted in this section. Of particular concern is the varying likelihood of *agreement* between the MTWs and the relative properties of the Tweets and their scores in each question in different decision scenarios.

The notion of score *disparity* is used to determine the difference in interest between a set of Tweets presenting with a range of different interestingness scores. To this end, each question asked of the MTWs has a disparity associated with it. The absolute score disparity, $d_G(q)$, for a given ranked question, $q \in Q$, is defined as:

$$d_G(q) = \max(s_G(t)) - \min(s_G(t)) \quad \forall \quad t \in q$$

Num. confident answers in q	min. $d_G(q)$	max. $d_G(q)$	avg. $d_G(q)$
0	0	846	**17.6**
> 0	0	1445	32.1
1	0	1445	**34.3**
> 1	0	4	0.647
> 2	0	0.55	0.204

Table 5.6: Illustrating trends between the absolute $d_G(q)$ with the varying number of confident answers made in q. Entries in bold are used to highlight interesting values.

Recall that a confident answer to a question is one where at least two of its three assessing MTWs select the same Tweet as interesting. Since an MTW could select more than one Tweet from each question, then each question may, in fact, have more than one confident answer. Table 5.6 illustrates how questions with varying score disparities can have an effect on the probability of MTWs being able to make a confident decision.

The results show that the average $d_G(q)$ of all $q \in Q$ is roughly double in cases where a question is answered with precisely one confident Tweet than in cases where there was no confident answer made at all, demonstrating how more interesting information is easier to identify when amongst noise.

If several Tweets with similar scores are listed, then it becomes more difficult for an agreement to be made between the different users on which Tweet is the *most* interesting. To reinforce this further, the average disparity was found to be much lower in cases where a question had multiple confident answers made. In these cases, the MTWs were unable to select one single Tweet as the most interesting and instead agreed on a set of top Tweets.

Let $d_{\text{sel}}(q)$ be defined as the disparity of the global scores of the MTW *selected* Tweets in question q. Table 5.7 highlights further the effect of disparity on human selection by demonstrating that, on average, $d_{\text{sel}}(q) < d_G(q) \; \forall \; q \in Q$.

This feature is particularly observable in cases where a question consists of a few Tweets having similarly high scores amongst Tweets with collectively lower scores. Therefore, inferring the interesting Tweets is easier, demonstrated by the scores of selected Tweets being generally higher, but discerning one *most* interesting Tweet is not as trivial. For example, the results show that, on average, $d_{\text{sel}}(q)$ is around 57% of $d_G(q) \; \forall \; q \in Q$.

	Average $d_G(q)$	Average $d_{\text{sel}}(q)$	Ratio
Average $s_G(t)$	62.4	35.3	0.57

Table 5.7: Comparing the average disparity of *selected* Tweets and the disparity of *all* of the Tweets in questions in Q.

5.7.4 Methodology and Validation Remarks

In this section, the proposed improvements to the inference methodology have been implemented and assessed under a randomised controlled trial using Amazon's Mechanical Turk to crowdsource the validations.

Results from the analyses indicate the method's relative advantages over the techniques used in the previous chapter. In particular, the new method is applicable to generating appropriate interestingness inferences for Tweets from all users on Twitter, is capable of effectively *ranking* Tweets in order of interestingness, and is far more efficient in model training. It is also much more readily supportive of 'on demand' inferences.

However, the crowdsourcing validations conducted were contributed to by people who shared no connection with the authors of the Tweets, and were thus assessing Tweets from outside of their own local network. As such, these evaluations are likely made on the basis of 'global interestingness', where Tweets that convey some meaning are highlighted from the noise, yet where the 'interesting' Tweets selected may not actually be relevant to the assessing MTWs. It is known, however, that Twitter users typically form followships between other users that produce information of both interest and relevance.

The following section addresses this by evaluating the accuracy of the scoring mechanism when users assess Tweets from within their own local network. This is of concern to the research in this Thesis since the users have a pre-defined interest in the authors of the Tweets they are evaluating.

5.8 Addressing Individual Information Relevance

In these analyses, results are studied from validations conducted through users assessing Tweets existing within their own local network. In particular, the interestingness scoring methodology will be validated against people's Tweet rankings for those users that they directly follow. Interactions with Twitter in this section relate to v1.1 of the Twitter API, as this research was conducted after the mandatory switch-over to this version.

Through assessment in this way, the Tweets being assessed are more relevant to their 'environments', which, in this case, consist of those users who would naturally also receive these Tweets and who are making the interestingness decisions based on their content.

5.8.1 Methodology

As will become clear, no initial data collection is required for these analyses. Instead, users contributing to the crowdsourced analysis interacted almost directly with Twitter during the course of their assessments, which involved the studying of Tweets sent from the friends of the assessing user.

For this purpose, a web application was set up in July 2013 and ran until August 2013, which allowed visiting users to 'sign in' using their Twitter account through OAuth. As with v1, v1.1 of Twitter's REST API directly supports this kind of behaviour, and provides the authenticated application with access keys enabling it to interface with the API on the authenticating user's behalf. Applications registered on Twitter can have different levels of access to users' accounts - from read-only, in which Tweets, follower information, and so on, can be retrieved; to read and write, with which new Tweets can be posted for the user and new followships can be created. An advantage of using OAuth in this fashion is that each user has a separate rate limit associated with it, meaning that the application could retrieve a lot of information, if necessary, yet without exceeding the rate limit afforded to application-only authentication by the new policies of v1.1 of the API.

In this case, the web application was advertised through word of mouth and through OSNs, such as Facebook and Twitter (see Figure 5.8) itself, as well as through Mechanical Turk. In the latter case, a special link to the site was provided to MTWs, and a code was displayed to them on completion of the task, which they could enter into Mechanical Turk in order to be paid. Participants contributing from the word of mouth and OSN categories are termed as ‘organic’ participants. Since the analysis depended on users assessing Tweets from their Twitter friends, participants could only take part if they had a Twitter account with at least 30 friends.

Figure 5.8: Advertising the validation site on Twitter.

After signing into the read-only application[5] and beginning the procedure, participants were faced with a series of ten Tweet timelines. The first, illustrated by Figure 5.9, consisted of the most recent 20 Tweets from the participant’s home timeline, and the next nine consisted of user timelines of the participant’s friends. Although the selection of friends for the nine user timelines was performed at random, a slight bias was applied towards selecting friends with a higher follower count through a weighted roulette-wheel selection. Due to the nature of scale-free graphs, there are many vertices with few edges, and few with many edges. As such, in order to obtain a more even distribution of user influence, the weighting was necessary to ensure that the scoring

[5]Twinterest: source available at https://github.com/flyingsparx/twinterestingness

mechanism could be validated against a range of users expressing Tweets with a wider variety of retweet counts.

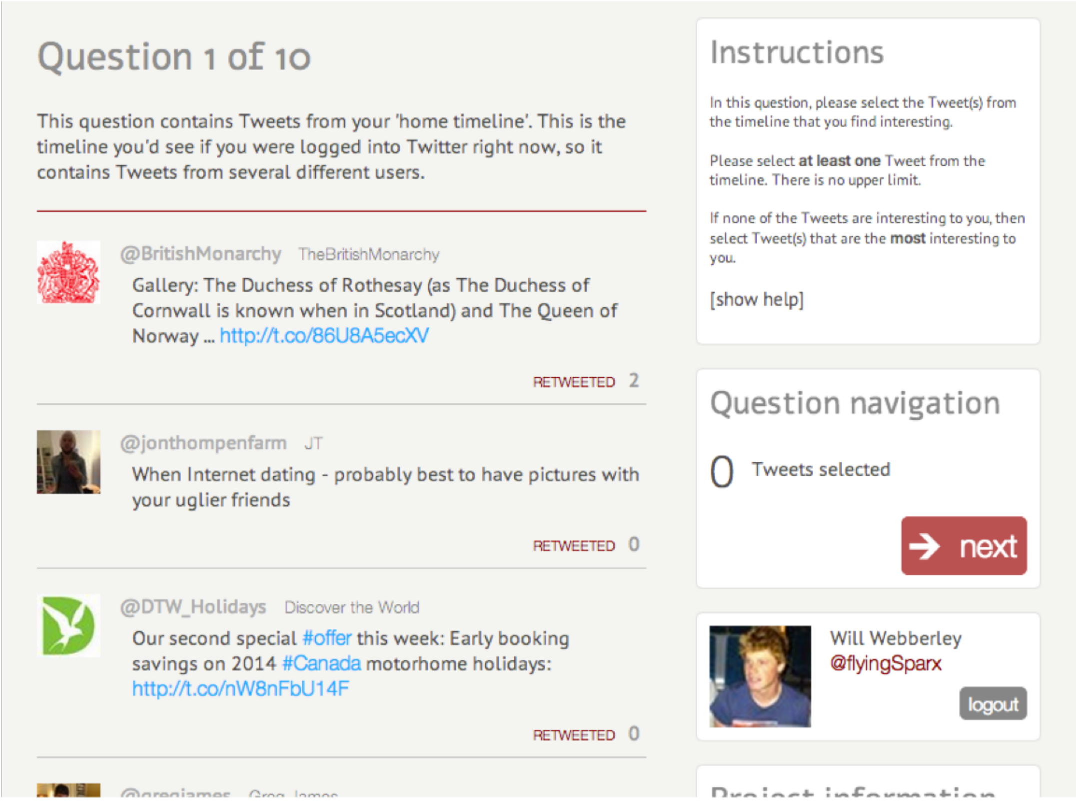


Figure 5.9: Screenshot of the experimental web application.

Participants were simply asked to select the Tweets that they found to be interesting from each of the timelines, as illustrated by Figure 5.9, and were not permitted to proceed to the next timeline without selecting at least one Tweet. Note that, at this stage, the Tweets being assessed did not have interestingness scores applied to them. A Tweet in a timeline that was selected was considered to be interesting, and those not selected were considered non-interesting.

5.8.2 Assigning Scores to the Assessed Tweets

A total of 580 timelines were assessed through the application validations, consisting of 389 contributed to by MTWs and 191 from organic participants. The totals are not precisely divisible by ten since not all participants assessed all of their ten timelines before leaving the application, but no single participant contributed more than ten timeline assessments. In this case, all responses were considered as confident since it was not appropriate under the conditions of the validation test to gain more than one assessment for each Tweet. Although there was likely some friend overlap between the participants, this was not necessarily the case in the vast majority of users assessed.

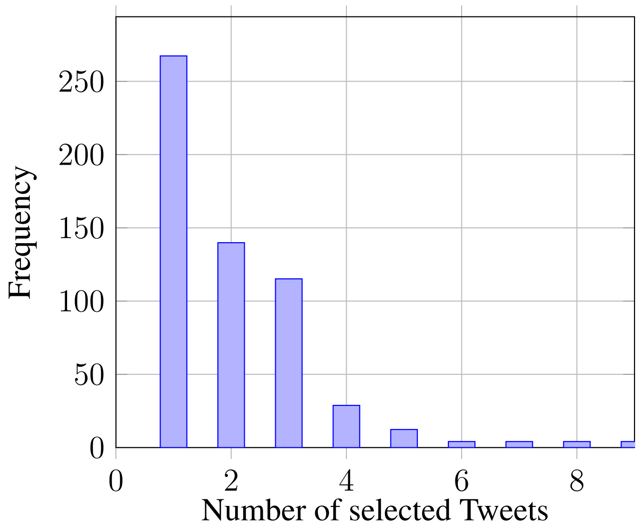


Figure 5.10: Frequency distribution of the number of Tweets selected from each timelines by the participants.

The validation test resulted in a set of just under 10,000 Tweets, authored by 936 unique users, that participants had made interestingness decisions on. Let T^b_{test} be the set containing these Tweets, and in order to determine their predicted expected retweet counts as part of assigning them scores, two procedures were required to take place;

- Collect further data on each assessed *author* in order to generate the 'user' models, and;
- Collect further data on each assessed *Tweet* in order to classify it against the previously-generated global model and the relevant user model.

The global model used was the same large model generated during the previous validation tests.

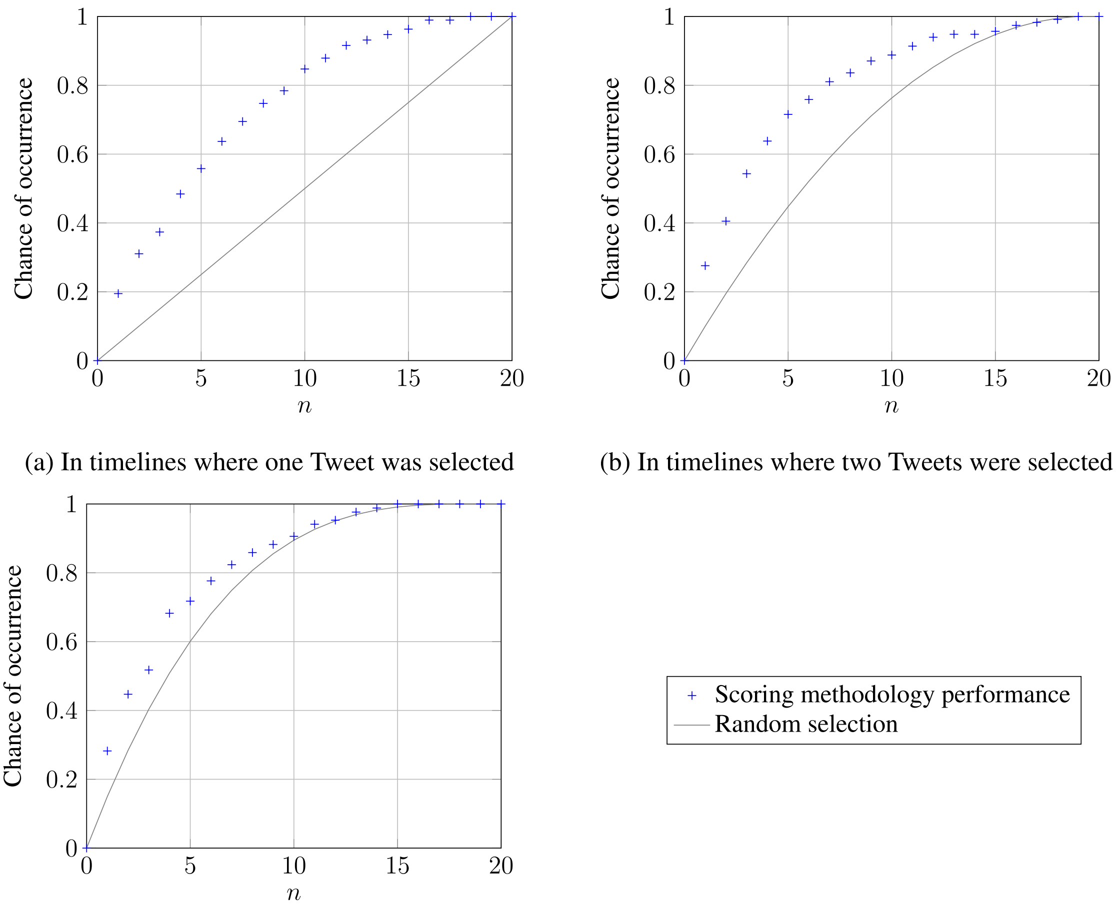
For reasons of privacy, each participant's Twitter API credentials were not maintained by the application and so standard authenticated REST API requests were performed to collect the additional data required. In particular, in August 2013, each of the 936 users $\{t.\text{author}_O \; \forall \; t \in T^b_{\text{test}}\}$ were queried under an identical collection scheme to that used as part of the previous validation; information on the author itself and on a sample of the author's followers and friends was retrieved. The collected information was also assigned to each of that user's Tweets $t' \in T^b_{\text{test}}$ so that an instance could be built for every $t \in T^b_{\text{test}}$ according to the features described in Table 5.5. These Tweets were then classified by the global model and their appropriate user model, which was built from its author's features, in order to eventually produce the two scores.

It should be noted that if a particular user follows another whose account is protected (see earlier in the thesis for further information on this), then the former user's API credentials can be used to view the latter's information and Tweets. However, since, during the data-collection, a static account was used to query the API, then Tweets and user information for accounts that are protected could not successfully be retrieved. This means that user and Tweet data for these users could not be collected for the purposes of training the user model and testing Tweets against this and the global model, and thus Tweets from protected authors had to be removed from T^b_{test}. The numbers stated in this section are those of the *final* dataset after removing these Tweets and users.

5.8.3 Results from the Further Validations

In this section, the patterns observed through the comparison of the Tweets inferred as interesting through the scores and those indicated as interesting by the human participants are analysed. The combination of both sets of participants was considered in the following analyses. As before, the $s_G(t)$ was used as the scoring scheme for the analysis in this section.

Performance of timeline ranking



(a) In timelines where one Tweet was selected

(b) In timelines where two Tweets were selected

(c) In timelines where three Tweets were selected

Figure 5.11: The chance of a participant selecting one of the *highest* n ranked Tweets in the timeline.

In the previous validation, the performance of the interestingness scores in ranking Tweets was assessed on a per-question basis. The same concept is expanded here to apply a similar assessment of the scores on the present validation test.

In this case, each assessed *timeline* was ranked in order of descending interestingness score in an effort to find the probability of a participant selecting a Tweet occurring in the top n of Tweets. Timelines were up to 20 Tweets long, compared to the five used in the Mechanical Turk questions in the initial validation test, but Figure 5.11 demonstrates that the mechanism is still able to effectively rank the Tweets. Since the timelines are larger than the questions used before, the chance of a participant selecting multiple Tweets from a timeline was greater, as indicated by Figure 5.10. To illustrate this, the results for this analysis are demonstrated against the appropriate random performance benchmark produced by the different selection criteria.

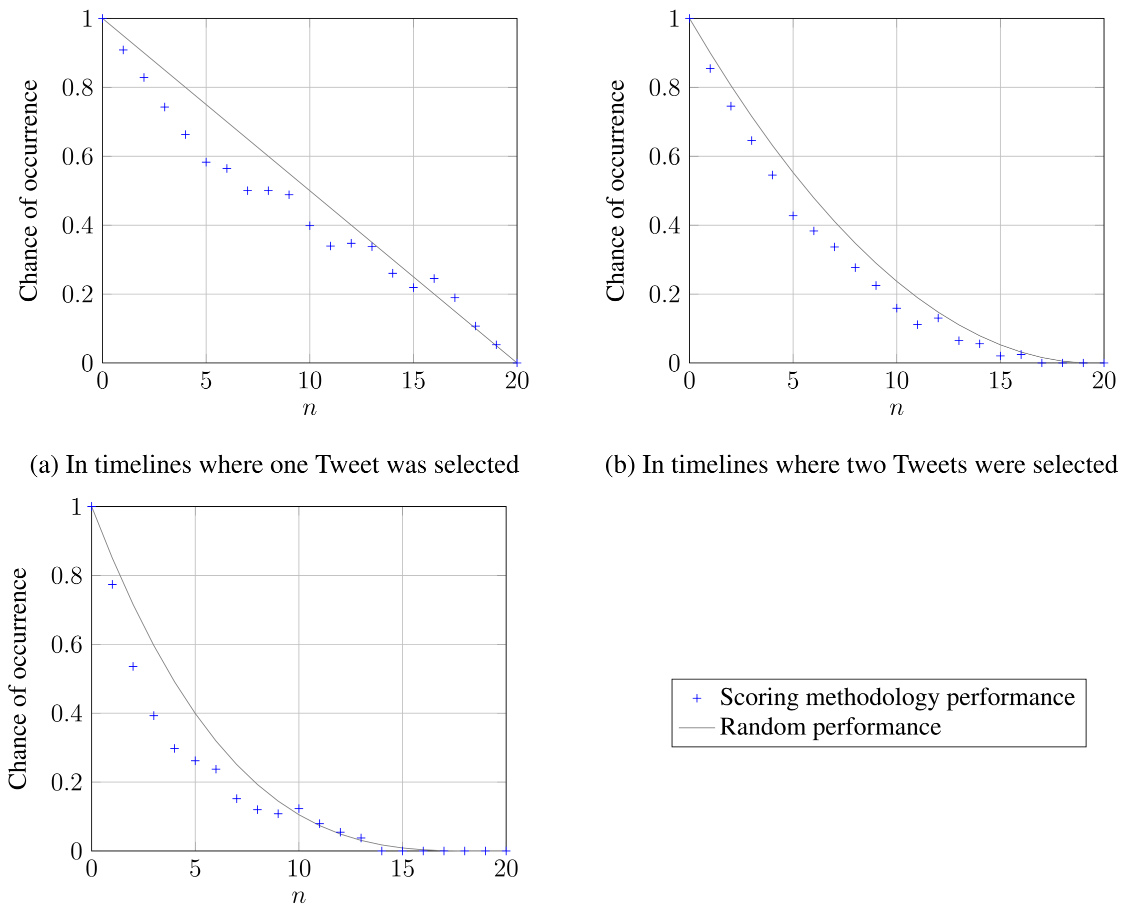
For Figure 5.11(a), the random performance is defined as the probability of a Tweet randomly selected from a timeline being in the top n of ranked Tweets in the timeline. In Figures 5.11(b) and 5.11(c), it represents the respective probabilities of the random selection of two and three Tweets being in the top n of ranked Tweets.

It is clear that the scores are able to identify interesting information from the noise around it, and so further analyses were conducted into the performance of the scores in detecting *un*-interesting information. In this scenario, each timeline had its Tweets ranked in order of *ascending* interestingness score and calculations were carried out into the probability of participants *not* selecting the *bottom* n interesting Tweets in each timeline.

Figure 5.12 illustrates this data, again with the different selection criteria. It is clear that although the scores are able to assist in identifying non-interesting information, the difference between this performance and the random selection case is not as great as with detecting the positively interesting Tweets.

Precision & Recall

Precision and recall have been useful ways of verifying the relative performance of the binning scheme for the simulation algorithm, in assessing the qualities of the various potential classifiers, and in several pieces of literature carrying out social network analysis, as highlighted in the Background chapter.



(a) In timelines where one Tweet was selected

(b) In timelines where two Tweets were selected

(c) In timelines where three Tweets were selected

Figure 5.12: The chance of a participant *not* selecting one of the *lowest* n ranked Tweets in the timeline.

The two metrics are also useful in demonstrating the performance of the scoring mechanism with respect to a varying *interestingness threshold*. Of concern is the similarity between the interestingness inferences made by the methodology and the interestingness decisions made by the participants. h is defined as an interestingness threshold, where a particular Tweet, t, is inferred as interesting only if $s_G(t) \geq h$. Precision and recall use a varying h value to determine the accuracy of the inference methodology, where;

$$\text{Precision} = \frac{\text{Of Tweets with score} \geq h\text{, number selected by participants}}{\text{Number of Tweets with score} \geq h}$$

$$\text{Recall} = \frac{\text{Of Tweets selected by participants, number with score} \geq h}{\text{Number of Tweets selected by participants}}$$

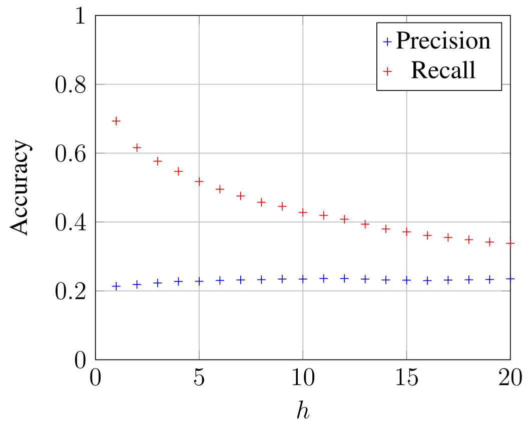


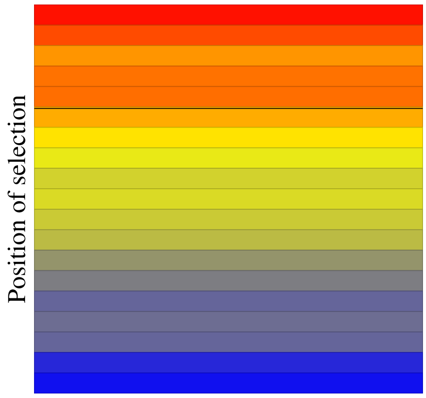
Figure 5.13: Accuracy (in terms of precision & recall) on the scoring mechanism with varying score threshold, h.

Figure 5.13 illustrates the precision and recall for T^b_{test} for $1 \geq h \geq 20$. As h increases, the number of Tweets with $s_G(t) \geq h$ decreases, causing the recall value to reduce. Although the precision increases slightly with greater h values (due to a smaller number of Tweets with higher scores), it is mostly consistent between 0.23 and 0.27 in the range shown.

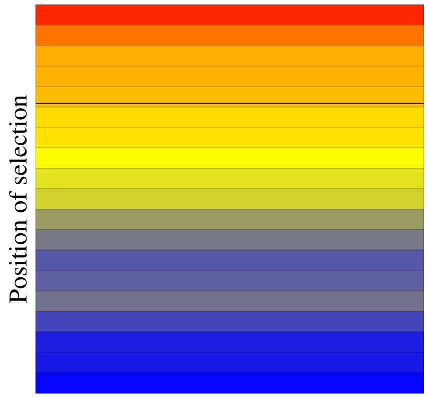
Since the recall is more variable, a value of $h < 5$ provides the best threshold for making a collective interestingness inference, as this provides a reasonable recall and precision in the context of the data.

Crowdsourced timeline selections

A brief study was additionally made into the selections of Tweets made by the participants. Of particular interest is the *difference* in performance between those participants that were paid to take part (the MTWs) and those who took part without being paid (the organic participants), and whether one group was more likely to select Tweets near the top of the timeline without scrolling down to read and select those at the bottom of the timeline. Reasons for this case could be laziness on the behalf of the participant, or simply for speed.



(a) Selections made by organic participants



(b) Selections made by MTWs

Figure 5.14: Heatmaps illustrating the timeline position of Tweet selections made by participants. Mean selection position is indicated.

Figure 5.14 shows the results of this study, and revealed that there was very little difference between the two participant groups. The organic participants, on average, selected the Tweet at position 6.07 in the timeline, and the MTWs selected Tweets at the average position of 5.83 out of 20 maximum available positions. Whilst these selection position averages are both relatively near to the top of the timeline, it should be noted that the *mean* timeline length was of 14 Tweets, and thus purely average random selections would be made at around the mark of the seventh Tweet.

It is felt, therefore, that there was some bias in both participant groups in that they were both slightly more likely to select Tweets nearer the top of the timeline than scroll down to view, and make interestingness judgments on those, nearer the bottom of the timelines. As with the previous validation tests, it was also possible to demonstrate that the score disparity between Tweets in a particular timeline is greater in cases where only one selection is made by the participants (Figure 5.15).

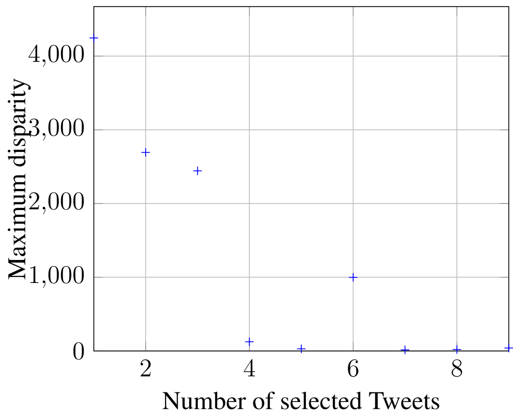


Figure 5.15: Relationship between the number of selected Tweets in a timeline and the maximum score disparity of the timeline.

5.9 Chapter Summary

In this chapter, an improvement over the previous iteration of the Tweet interestingness inference methodology has been introduced, tested, and analysed. Bringing the research into the social structure of Twitter forward from the previous chapter, it was possible to determine areas for improvement and the useful metrics for governing the selection of new features.

Question **RQ4** from the initial hypotheses has been answered in order to show that interestingness of Tweets can be inferred non-semantically with some degree of accuracy, and the methods have been able to demonstrate their ability to rank Tweets in order of interest. It is clear how the method could be used for highlighting particularly outstanding Tweets in order to, for example, identify the more controversial Tweets surrounding a particular event, and for producing a digest of Tweet 'timelines' based on estimated interest rather than time.

5.9.1 Interestingness Scores

The new methodology introduced the notion of scores, which can be assigned to Tweets in order to signify their relative interestingness. These scores are based on the ratio between the popularity of a Tweet, measured by its observed retweet count, and a value representing an *expected* retweet count for the Tweet. The scoring mechanism works such that different types of users, including those ranging across influence and activity frequency[6] levels, can have their Tweets assessed on the same scale.

Two scoring schemes were set up, which are derived from distinct methods for generating the expected retweet count. One method is based on comparing a Tweet's (and its author's) features to a *global* model trained on a large number of Tweets collected from Twitter. The other is generated through the comparison of the Tweet's features to a *user* model trained only on other Tweets posted by that same particular user. Generally, there was found to be a non-significant difference between the performance of the two scores, however, and thus they were both used interchangeably during the validations.

5.9.2 Methodology Validations

Two sets of validations were conducted into verifying the performance and accuracy of the new methodology and the scores it produced - one in which Tweets were placed into questions on Amazon's Mechanical Turk, in which MTWs were asked to select the most interesting Tweets; and another, in which participants were asked to sign-in through Twitter and then assess Tweets from users they actually follow.

In the first case, the participants shared no connection with the authors of the Tweets they were assessing (except in the case of coincidences), and were therefore assessing Tweets on a *global* interest level. In particular, this largely involves determining the

[6]The activity frequency is the rate at which a user posts Tweets.

interesting information from the noise around it. In the second set, information *relevance* came more into play, since participants were assessing Tweets from users they have already declared an interest in (through the action of following).

In both test cases, the validations showed the scores to be able to appropriately label Tweets according to interestingness in a variety of different ways. The second test included an analysis demonstrating that the scores are more efficient at determining interesting Tweets than *un*-interesting Tweets, the latter of which would be useful in deciding on a set of Tweets to discard from an interesting set.

5.9.3 Improvements and Qualities

The newly introduced methodology presents several improvements over that described in the previous chapter. In particular, the performance of the scores have shown a large accuracy improvement in determining interesting information. The previous method also did not take into account *how* interesting a particular Tweet may be, and was only able to make a binary interesting/uninteresting decision for each Tweet.

Another large improvement is the ability of the scoring method to be applied to a much wider range of Tweets. The previous method was realistically unable to assess Tweets from users with more than 300 or so followers due to data collection inefficiency, the time taken, and the computational complexity involved in simulating large user graphs. The new method can be used to assign scores to Tweets in a more "on demand" fashion, where only a small amount of information for each Tweet is required in order to generate the features needed to predict the estimated retweet counts. The scores also allow Tweets from many different sources to be assessed on the same scoring scale, meaning that Tweets on a mixed timeline can be appropriately compared to one-another, as demonstrated by the second set of validations.

Chapter 6

Assessment and Conclusions

Here is provided an overview and assessment of the work conducted in this thesis, bringing together the ideas from the initial research and how these have helped in developing the methodologies introduced in later chapters. The validations from the methods are further assessed, followed by an explanation of how the research forming them may be taken further in potential future projects. Finally, an overview of the thesis in terms of its contributions is described.

6.1 Analysis of Research and Results

In this thesis, the research behind the development of an effective interestingness predictor has been described. The relative successes of the methodology and its advantages over its previous iterations have been illustrated through in-depth analyses of its performance in various situations.

Following is an analysis of the research carried out over the main stages described in the primary chapters of this thesis.

6.1.1 Retweeting & the Twitter Structure

Initial research was conducted into the act of retweeting and Twitter in general for the purposes of providing a background and foundation for the later work. In particular, the properties of retweet groups and the behaviour of the users within them was demonstrated. During the review of the relevant literature in the field, it was suggested that a Tweet's popularity cannot generally be directly tied to the Tweet's interestingness due to factors relating to user *influence*.

Agreeing with other research in the area, it was demonstrated that retweet groups can have widely ranging sizes and depth. This observation takes into account that retweets can, themselves, be retweeted, and that retweet groups do *not* consider the followships between the set of users they represent. Retweet groups were found to present an average *maximum* path-length of around two, and the longest maximum path found in the dataset collected from retweets on the public timeline was of length nine. This demonstrates a significant penetration through the social graph, especially considering the 'real' world's six degrees of separation, and that social networks often exhibit a social graph even more closely connected than this.

It was found that the chance of a retweet occurring was much greater in cases where the retweeter follows the author of the original Tweet. As retweet pathways become longer, the chances of the final retweeter following the original author diminishes over the distance, demonstrating strong correlations between the edges separating users on the retweet graph and those on the social graph. These experiments were conducted using a trained logistic regression to predict a retweet outcome decision for each user who received a particular Tweet during simulations of Tweets through each structure type.

The correlations and results from the explorative analyses on the social structure and the arrangement of users on the social graph indicated that the social structure of Twitter clearly affects the propagation of retweets and that this property could provide a useful way of estimating Tweet interestingness. This triggered research focussed on examining the differences in propagation patterns in order to demonstrate that each structure type can present very different retweet propagation patterns. Because the propagation pattern difference at this structural level was so large, it was decided that this could be a basis for an interestingness inference methodology. This method utilised the same research and algorithms behind those used in the graph structure analysis to predict a retweet count for a given Tweet within a graph of connected users, and worked through a simple comparison between this predicted value and the *observed* retweet count of the Tweet. This method was not shown to perform particularly well in the validation tests conducted, and thus improvements were necessary before any further analyses were made.

6.1.2 Interestingness Scores

Improvements over the previous methodology were based around the introduction of interestingness *scores*, with which Tweets could be ranked according to the ratio of their observed and expected popularities, and where if the observed popularity is proportionately larger than the expected popularity, the score for that Tweet would also be proportionately greater. This in itself provides many benefits over the previous system, which was unable to provide any indication over *how* interesting a Tweet is.

The prediction of the estimated retweet count was altered so that they could be generated directly through the use of a Bayesian network machine learning classifier, which made predictions based on a larger set of Tweet and environmental network features. These features could be collected much more efficiently from Twitter's REST API, illustrating another advantage in terms of the ease with which predictions (and thus score assignations) can be made.

Furthermore, the efficiency stretches to providing a more universal approach, allowing Tweets from most users on Twitter to be evaluated equally and on the same scale, since the complexity of any part of the assignation process is not affected by the influence or other properties of the author user.

A Bayesian Network was chosen for the methodology improvements due to its relative advantages over the other assessed classifiers, as highlighted by Table 5.3. Although the logistic regression performed almost as well as the Bayesian Network in the cross-validations, its training time, especially with the full dataset, meant that it would be unsuitable, especially if user models are to be generated on-demand. Logistic regression was more appropriate in the earlier work described in Chapter 4, in which of concern was its ability to produce a retweet *probability* from a set of binary feature values.

6.1.3 Validation

Producing the scores partially relies on the initial accuracy of making retweet predictions in the *general case*, using cross-validation tests on the Bayesian network classifier and the binning policy of retweet counts explained in the previous chapter. The performance of each factor, and the accuracies achieved, are highlighted in Tables 5.3 and 5.4 respectively.

The mention of the "general case" is important, since the methodology is designed to discover Tweets which do *not* fit this case, as these would be the Tweets which have a greater (or smaller) retweet count than expected, and would therefore be the Tweets which would contribute negatively to the aforementioned performance analyses of the prediction method. As such, if all Tweets fit their general cases as given by their features and the features of their authors, then the general performance of the cross-validations on the classifier could be greater, but then no interestingness inferences could be made.

Two main human validation tests were conducted into the performance of the scoring mechanisms provided by the improved methods; one based on interestingness decisions from non-related participants, and another based on decisions from Twitter users to whom the Tweets assessed are more directly relevant, as denoted by the followships of the author users. These validations expressed a good performance of the scoring scheme in a variety of ways, from the ranking of Tweets in order of interestingness through to analyses into the motivation of Tweet selection from the *disparity* of Tweet scores in the timeline.

6.1.4 Methodology Evaluation

The Background chapter of this thesis described other similar research in this area along with the strengths and weaknesses of each. Whilst this included research into retweet decision and count predictions, they are often quite similar to one another, and these goals are not the primary focus of the work in this thesis. Instead, research into information interestingness with regards to Twitter will now be evaluated against the methods outlined in this thesis.

Gransee et al. [25] introduced a system for scoring Tweets based on a naïve Bayesian classifier. The authors' learner was concerned only with textual cues for producing a score, and thus the method is based on semantics. The learner was trained using a set of Tweets from a particular author, with each Tweet being assigned a score based, similar to the work in this thesis, on the distance between the observed retweet count of the Tweet and the single *baseline* retweet count for the Tweet's author at that particular time. Words in unseen Tweets are then scored individually according to the scores of Tweets the words have previously been seen in, which, when averaged, generates a score for the unseen Tweet.

The scoring method discussed in this thesis is superior to the methodology described by Gransee et al. [25] in a number of ways. Firstly, the method requires a pre-built dictionary of words to be generated and scored for each user before a model can be trained and any scores can be assigned, meaning that it is not possible to carry out on-demand assignations of scores to Tweets. Additionally, a baseline retweet count needs to be maintained for each user assessed for specific time-intervals, causing the necessity of periodically updating the word dictionary in order to reflect the change in baseline. The authors admit that their methods work better when Twitter users are more predictable and use similar words across Tweets. The global model discussed in this thesis is able to represent snapshots of projected Tweet popularity for many types of users, and is therefore re-usable for a single user as he/she gains or loses influence, and does not require regeneration. Finally, assessments were only made for the *top* 175 users on Twitter, with no indication of its performance on users who are less influential and typically receive far fewer retweets per Tweet and users who do not Tweet enough in each time-frame for a suitable baseline value to be calculated. Influential users are generally subject to smaller fluctuations in the social graph, as additional followers wouldn't have such a large impact, and thus the retweet count baseline wouldn't vary as often as with less influential users. Therefore it would be more difficult to apply this method to Tweets from 'normal' (or less influential) users.

The semantic scoring of Tweets was also considered by Alonso et al. [4], but with a focus on determining non-interesting information. In this case, the authors used a scoring scheme that assigned an integer value to each Tweet of between 0 and 5. However, their results are largely based on simply marking a Tweet as interesting if it contains a URL, which, again, does not allow the method to be appropriately applied to a wider range of Tweets. Crowdsourcing is used as part of obtaining human input for determining interestingness, but the MTWs are instructed on what to class as interesting instead of allowing participants more reign on what constitutes the *most* interesting. Generally, these features are not comparable to those in the methods discussed in this thesis, since the authors are more concerned with finding *un*interesting Tweets to be disregarded from a stream, have a limited scoring range and do not conduct as rigorous validations into their results. Despite achieving some accurate results, the methods cannot be used on such a wide variety of Tweets as is available on Twitter.

Finally, Lauw et al. [36] presented a study that used a clustering algorithm to estimate a Tweet's *quality* based on a function of a Tweet's audience size and its relation to other similar Tweets. The authors' research is based around news stories, in that Tweets 'belonging' to the same event are clustered together and are assigned a quality depending on the size and 'importance' of the cluster. Despite the similarities in terms of assigning a quality to clusters and Tweets, the authors do not take into consideration many of the factors discussed in this thesis (including the notion of timeline 'slip' discussed in Chapter 4, which is relevant to their work) and the verifications conducted into the performance of their methods are not rigorous enough to prove that the quality level assigned to each cluster or Tweet is accurate or appropriate.

Here (Table 6.1) is provided an overview of the key results from the validations of the scoring method.

Result	Description
Non-semantic **identification** of globally interesting information	Section 5.7 demonstrates the method's ability to determine globally interesting information from a set by indicating that Tweets with higher scores are more likely to be selected as interesting. The method involves no semantic understanding of the Tweet content.
Ranking globally interesting information	Section 5.7 demonstrates the method's ability to rank information by interestingness. For example, participants agreed that one of the two top score-ranked Tweets were interesting in groups of five in 66% of cases.
Addressing information **relevance**	Section 5.8 involved validations of Tweets 'relevant' to users and shows that the method is able to rank longer timelines of Tweets. For example, participants agreed that one of the top five Tweets were interesting from timelines of 20 in 57% of cases.
Identifying **non-interesting** information	Results in Section 5.8 show that the method is not as appropriate for identifying non-interesting Tweets as it is for determining interesting Tweets.
Effect on **precision & recall**	Raising the score threshold that 'define' interestingness does not have a large effect on precision but does reduce recall considerably.

Table 6.1: Overview of key results.

6.1.5 Contributions

Throughout the earlier chapters of this thesis, work has been conducted towards answering the hypothetical questions asked in Section 1.3 in the Introduction. The research has highlighted the inappropriateness of using retweet counts alone in indicating interestingness, and has shown the impact of the layout of users on the social graph on message propagation and how this can be useful for estimating Tweet interestingness. In particular, the questions are now answered more formally.

RQ1 *- Does Tweet popularity, measured in terms of retweets,* ***define*** *interesting (or non-'noisy') information?*
The number of retweets a particular Tweet receives cannot appropriately be used for defining interesting information. A Tweet that has been retweeted at least once is not necessarily interesting.

RQ2 *- Can Tweet popularity, measured in terms of retweets,* ***be an indicator*** *of interesting information?*
The number of retweets a Tweet has achieved is not alone indicative of its level of interestingness. The overall retweet count of a Tweet is produced as a function of its author's influence, and therefore Tweets written by different authors cannot have their Tweets' interestingness measured on the same scale.

RQ3 *- Is the arrangement of Twitter's social graph an important factor in retweet propagation, and thus perceived popularity?*
The layout of users and the edges connecting them on the social graph has been shown to strongly affect the permitted propagation of Tweets; some structures facilitate retweet spread, whilst others throttle it. The edge density also partially dictates users' influence levels, in that those users who are assigned a larger in-degree are more likely to achieve more retweets due to the magnitude of spread of the original Tweets. This importance was brought forward to the later work in identifying interesting Tweets.

RQ4 *- Can Tweet interestingness be inferred* ***non-semantically****?*

Chapter 5 showed that a Tweet's interestingness can be determined through a study of the Tweet's features and those of its author and the latter's relationship to others on the social graph. The inferences are therefore non-semantic, since they do not make any attempt to understand the content of individual words or phrases in the Tweets' contents. The method was also demonstrated to be suitable for ranking Tweets in order of interestingness *level*, so that Tweets estimated to be more interesting could be shown at a higher priority.

As part of the research, several contributions to social media analytical research have been made, which have been discussed in the relevant chapters of the thesis and are summarised below.

A comprehensive survey is carried out into relevant literature in information propagation in online social networks, along with evaluations and assessments of some of the research more specific to inferring interestingness.

Retweet properties and the way they are influenced by the social graph are thoroughly researched in order to provide a general background and understanding of the notions relating to propagation in online social networks, propagation and information penetration, Tweet 'audience', and how these factors are related to the arrangement of users on the social graph. The definition of terms, such as 'retweet group' and 'maximum path-length' are useful for discussing various properties relating to retweeting on Twitter.

An investigative study is made into the use of machine learning techniques and classifiers in the field of social media, and the ways in which they can be useful for different purposes.

Finally, a numerical 'definition' (or quantification) of estimated interestingness is contributed. A method for suitably predicting estimated retweet counts as dynamically nominal categories is described, implemented and verified, leading to the development of a method for assigning interestingness scores to Tweets as part of an effort to support the ranking of Tweets and of highlighting interesting Tweets from the noise of everyday communication on Twitter.

6.2 Limitations

Although the Tweet-scoring method has the strengths identified above, it is also compromised.

In its current form, the technique can only be used on non-live Tweets that exist as part of a 'static' dataset. This limitation stems from the necessity of target Tweets having a representative observed retweet count associated with them. Tweets, on average, do see the significant part of their overall retweet action occur within the first hour to the first day, but this does mean that Tweets cannot be scored and suggested as interesting as they arrive on a timeline. This problem could be partially addressed by developing a model to extrapolate retweet *rates* for Tweets, which could be used in conjunction with the present model in order to predict a 'final' retweet count.

Even on static data sets, however, the accuracy is not particularly high. The results from the evaluation do indicate that identification of interesting Tweets is possible through the method, and that the ranking of Tweets based on the score is more useful than a random ordering, but the precision (as indicated in Figure 5.13) is quite low. Although the method is designed to be applicable over all Tweets and users and covering the *global* interestingness of information (including that which might not be relevant), improvements could be made so that the accuracy increases with the thresold used for defining interestingness.

Figure 5.12 represents the results of evaluating the scoring method's ability to rank Tweets in ascending order, by assessing how often scored-uninteresting Tweets *weren't* selected by participants. This performance is considerably lower than the inverse (likelihood of selecting scored-interesting Tweets), meaning that the identification of non-interesting or noisy Tweets is less successful. Improvements to the predictive model would provide more scope for the method to filter out Tweets that are less useful for people to read.

In the validations, the predictive model was evaluated against a dataset created by human participants. This set is a *gold standard*, since it represents the 'ground truth' on which Tweets are interesting and which are non-interesting. The human participants bridge the gulf between the semantics of Tweet contents and the non-semantic features of the Tweet, its author, and its author's network features. These models are incompatible since the humans cannot incorporate the underlying non-semantic features into their semantic decision-making process when creating the gold standard dataset.

6.3 Further and Future Work

There are various ways in which the work and research in this thesis could be taken further in extended work and projects.

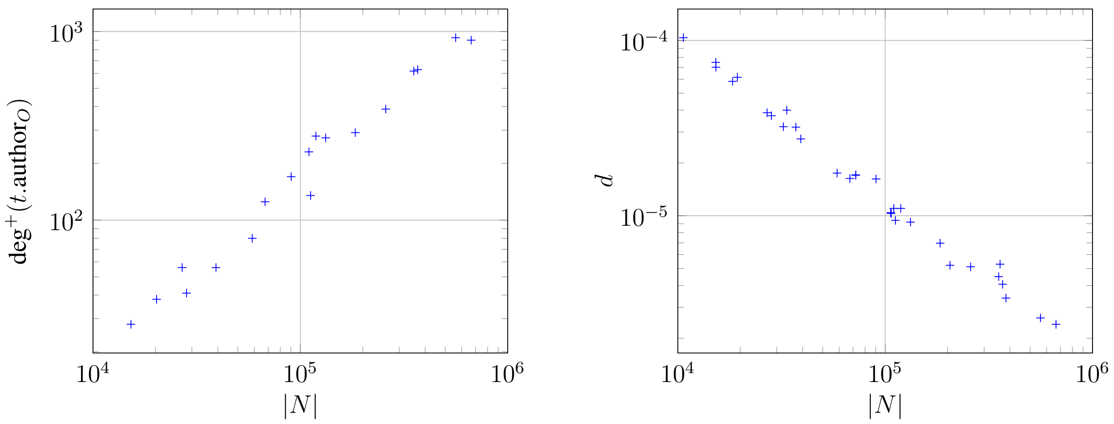
6.3.1 Building on the Social Structure

Chapter 4 of this thesis included research into the social structure of Twitter and the role this played in the retweet propagation patterns of the information flow within it. The impact of the structural changes on the propagation pattern led the way in building the initial interestingness inference methodology, which was based on simulating Tweets and retweet decisions through social structures reflected directly from Twitter.

The method had many drawbacks, as described earlier, but mainly these were due to its inefficiency in terms of the amount of data necessary in order to construct the local networks, which included all the users and edges within two hops of the source author user. Nodes three hops or more from the source were prohibitively complex to collect and implement due to the scaling nature of the Twitter social graph, and two was considered sufficient as it was the average *maximum* path-length of Tweets analysed in

Chapter 3. Despite this, the limitations in data collection meant that Tweets could only realistically be simulated for users with much smaller local networks, and thus the methods could not be applied to a wide variety of Tweets on Twitter or have inferences made on demand.

The research in this thesis then followed a different path in an attempt to solve the presented problems, but there are ways in which the previous method could be improved more directly. One way, in particular, would be to attempt to *infer* or 'estimate' a user's local network from a set of its immediately-available parameters. Analytical results from this first methodology highlighted correlations between a user's follower count and the size of the user's local follower network, and a correlation between the size of this local network and the edge density of the users within it, as illustrated by Figures 6.1a and 6.1b.



(a) Relationship between local network size ($|N|$) and the follower count of an author of Tweet t.

(b) Relationship between local network size ($|N|$) and local graph edge density (d).

Figure 6.1: Plots illustrating indirect correlations between an author user's follower count and the edge density of the author's local network.

A graph's edge density can be calculated as a function of the number of nodes and edges existing within the graph;

$$d = \frac{|E|}{|N|(|N|-1)}$$

This suggests that a user's follower count could be used to indirectly estimate the edge density of that user's local follower graph. Since Figure 4.5 showed the similarities, along with other research covered in the Background chapter, between the structure of Twitter's social graph and that of a similarly-sized scale-free network, it is indicative that a faux scale-free network could be generated with a provided edge density and size in such a way as to reflect the user's own local network. Although this would not be an intricate mapping of the users and edges existent in reality, it would provide an appropriate size and density of edges between the nodes, within which Tweets could be propagated and examined.

Through generating a graph structure in this fashion, then there is no need to collect the local network information for each user to be assessed, since a faux graph could simply be created on the fly, and Tweets from more influential users could also be simulated (depending on the computational complexity of the generation algorithm and simulator). The *accuracy* of the interestingness inference could then be improved through a larger training set, the use of more features for handling the decisions, or by introducing further routines to make the simulations and the decisions more realistic.

6.3.2 Taking the Scoring Methodology Further

There are various ways in which research could be carried out into the methodologies behind the scoring mechanism used in the interestingness inferrer discussed in Chapter 5. In particular, one could be to research into 'unnecessary' edges between users in OSNs. The final validations conducted on this method involved participants assessing the Tweets they'd naturally receive onto their home timeline as they are from users that they already *follow*.

However, as has been made clear at points in this thesis, not *all* the information a friend posts is likely to be interesting, and that particularly interesting information from a particular source is likely to be retweeted more than less interesting information posted by the same source.

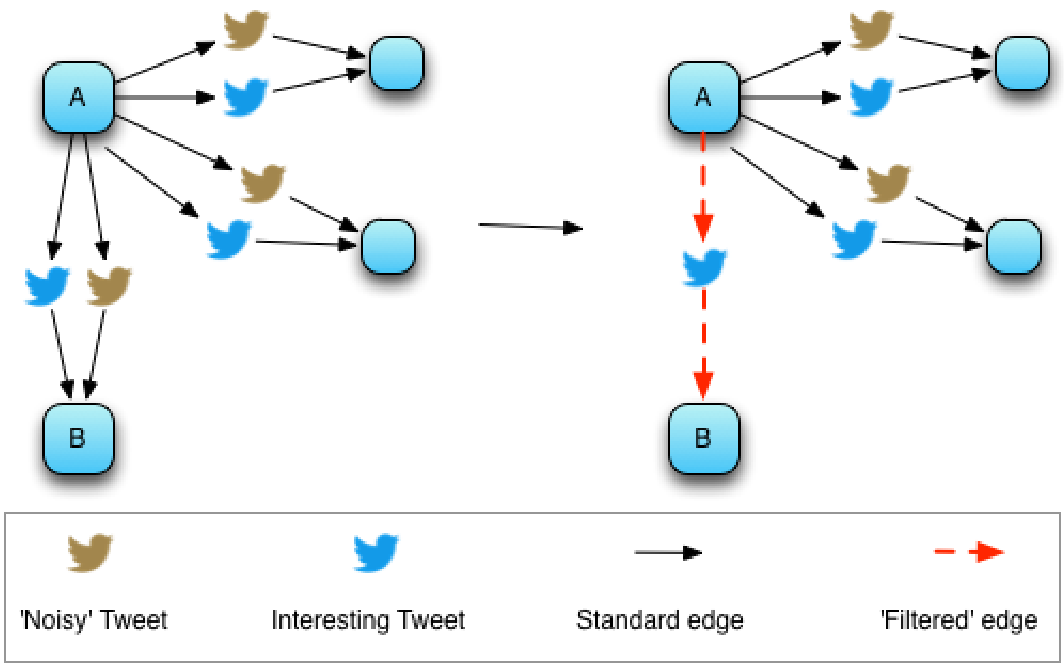


Figure 6.2: Using 'filtered' edges to improve interestingness precision.

Therefore, the interestingness scores could be used to allow users to form conditional followships to other users, down which only interesting Tweets are sent. Figure 6.2 initially shows user B receiving all of the Tweets from user A down a normal Twitter followship edge. However, if user B instead followed user A through a 'filtered' edge, then only Tweets with a sufficient interestingness score would be received. Under this type of scheme, user B avoids receiving the uninteresting Tweets from A, thus increasing the chance of noticing the interesting ones.

Of course, this system does require a set of users to be already following A in order to initially generate scores for each Tweet, but it does illustrate an interesting path for future research. Whilst largely hypothetical at this stage, it would provide an interesting extension to the research carried out in this thesis, in that interestingness scores could be used to assign thresholds of interestingness to users with the aim of discovering whether interesting Tweets could still be propagated to the extent that they deserve, yet by simultaneously reducing the propagation of noisy information.

Since it has been explained how a user's followships act as a 'search term' for information retrieval on Twitter, by finding ways of forwarding interesting information to users who do not directly (or 'traditionally') follow a source, then it is clear how this could pave a way for enabling interesting and *relevant* information can be delivered to users, yet without them having to look for it or know about its existence in the first place.

6.4 Final Remarks

The work in this thesis was carried out with the aim of researching a methodology that is able to suitably infer interesting information in Twitter. Although the research has focused on Twitter as a platform for information dissemination, the motivation for this research stems from the *noise* observed every day in all online social networks that support information propagation, including those such as Facebook and Tumblr.

The research processes culminated in the development of a method that ranks Tweets by assigning a score indicating an estimated level of interestingness based on a function of its perceived popularity. The methods were verified through the use of crowd-sourced validation tests covering the notions of general interestingness (in terms of identifying it from 'noisy' information) and of more *relevant* interestingness (through assessments of Tweets authored by more relevant users).

Analyses into the validation tests demonstrated the process by which users are able to identify interesting information and showed that the scoring mechanism is able to effectively rank Tweets in an appropriate order of interestingness in both mixed timelines and in timelines of Tweets authored by the same user. The scores can be applied to Tweets from all users on the same scale, meaning that inferences are not limited to a specific subset or type of Tweet.

The methods are open enough and use resources that are mostly common to many similar social networks. For example, "shares" and "reblogs" can be examined as the propagation mechanisms in Facebook and Tumblr, respectively, in attempts to apply the same scoring schemes to other platforms.

The research work represented by this thesis is being developed further towards applications in other projects. Specifically, it is being used to identify the more outstanding and controversial Tweets posted by particular users in research projects involved with police. Using a pre-trained model means the method does not need to be tailored for any particular use-case, and the application's necessity to produce information on-demand as part of a conversational interface makes the method useful and appropriate.

Bibliography

[1] Arifah C. Alhadi, Thomas Gottron, Jérôme Kunegis, and Nasir Naveed. LiveTweet: Microblog Retrieval Based on Interestingness and an Adaptation of the Vector Space Model. In *Proceedings of the 2011 Text REtrieval Conference*, TREC'11. National Institute of Standards and Technology (NIST), 2011.

[2] Arifah C. Alhadi, Thomas Gottron, Jérôme Kunegis, and Nasir Naveed. LiveTweet: Monitoring and Predicting Interesting Microblog Posts. In *Advances in Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 569–570. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-28996-5.

[3] Stuart M. Allen, Gualtiero Colombo, and Roger M. Whitaker. Uttering: Social Micro-Blogging Without the Internet. In *Proceedings of the Second International Workshop on Mobile Opportunistic Networking*, MobiOpp '10, pages 58–64, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-925-1.

[4] Omar Alonso, Chad Carson, David Gerster, Xiang Ji, and Shubha U. Nabar. Detecting Uninteresting Content in Text Streams. In *Proceedings of the 2010 Workshop on Crowdsourcing for Search Evaluation*, CSE'10, pages 39–42. ACM, 2010.

[5] Isabel Anger and Christian Kittl. Measuring Influence on Twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, page 31. ACM, 2011.

[6] Paavo Arvola, Jaana Kekäläinen, and Marko Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13:460–484, 2010.

[7] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four Degrees of Separation. *Computing Research Repository (CoRR)*, abs/1111.4570:33–42, 2011.

[8] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Identifying 'Influencers' on Twitter. In *Fourth ACM International Conference on Web Seach and Data Mining*, WSDM 2011. ACM, 2011.

[9] Alexander Balinsky, Helen Balinsky, and Steven Simske. On the Helmholtz Principle for Data Mining. In *Proceedings of 2011 International Conference on Knowledge Discovery, Chengdu, China*, pages 99–102. IEEE, 2011.

[10] Carolina Bigonha, Thiago N. C. Cardoso, Mirella M. Moro, Virgílio A. F. Almeida, and Marcos A. Gonçalves. Detecting Evangelists and Detractors on Twitter. In *18th Brazilian Symposium on Multimedia and the Web*, WebMedia, pages 107–114, 2010.

[11] Alex Burns and Ben Eltham. Twitter Free Iran: an Evaluation of Twitter's Role in Public Diplomacy and Information Operations in Iran's 2009 Election Crisis. In *Communications Policy & Research Forum 2009*, pages 322–334, November 2009.

[12] Guido Caldarelli. *Scale-Free Networks: Complex Webs in Nature and Technology*. OUP Catalogue. Oxford University Press, 2007.

[13] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4.

[14] H. Burak Celebi and Susan Uskudarli. Content Based Microblogger Recommendation. In *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and of the 2012 International Confernece on Social Computing*, PASSAT, SocialCom, pages 605–610. IEEE, 2012.

[15] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM*, 10: 10–17, 2010.

[16] Martin J. Chorley, Gualtiero B. Colombo, Stuart M. Allen, and Roger M. Whitaker. Better the Tweeter you Know: Social Signals on Twitter. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International*

Confernece on Social Computing, PASSAT, SocialCom, pages 277–282. IEEE, 2012.

[17] Aaron Clauset, Cosma R. Shalizi, and Mark E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM review*, 51(4):661–703, 2007.

[18] Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976. ISBN 0192860925.

[19] Chad Edwards, Patric R. Spence, Christina J. Gentile, America Edwards, and Autumn Edwards. How Much Klout Do you Have? A Test of System Generated Cues on Source Credibility. *Computers in Human Behavior*, 29(5):A12 – A16, 2013. ISSN 0747-5632.

[20] Paul Erdős and Alfréd Rényi. On the Evolution of Random Graphs. *Publications of tke Matkemafical Insfifufe of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[21] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian Network Classifiers. *Machine Learning*, 29(2-3):131–163, 1997. ISSN 0885-6125.

[22] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In *Proceedings of the 3rd Conference on Online Social Networks*, WOSN'10, Berkeley, CA, USA, 2010. USENIX Association.

[23] Liqiang Geng and Howard J. Hamilton. Interestingness Measures for Data Mining: A Survey. *ACM Comput. Surv.*, 38(3), Sep 2006. ISSN 0360-0300.

[24] Daniel G. Goldstein and Gerd Gigerenzer. *The Recognition Heuristic: How Ignorance Makes us Smart*. Oxford University Press, 1999.

[25] Sean Gransee, Ryan McAfee, and Alex Wilson. Twitter Retweet Prediction. 2012. URL `sites.google.com/site/learningtweetvalue/other-deliverables`.

[26] Oliver Hein, Michael Schwind, and Wolfgang König. Scale-Free Networks. *Wirtschaftsinformatik*, 48(4):267–275, 2006. ISSN 0937-6429.

[27] Suzanne Hidi and William Baird. Interestingness â A Neglected Variable in Discourse Processing. *Cognitive Science*, 10(2):179–194, 1986.

[28] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting Popular Messages in Twitter. In *Proceedings of the 20th International Conference companion on World Wide Web*, WWW '11, pages 57–58, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0637-9.

[29] David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. Wiley, 3 edition, 2013. ISBN 978-1-11854-839-4.

[30] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social Networks That Matter: Twitter Under the Microscope. *First Monday*, 14, 2009.

[31] Farhad Hussain, Huan Liu, Einoshin Suzuki, and Hongjun Lu. Exception Rule Mining with a Relative Interestingness Measure. In Takao Terano, Huan Liu, and ArbeeL.P. Chen, editors, *Knowledge Discovery and Data Mining. Current Issues and New Applications*, volume 1805 of *Lecture Notes in Computer Science*, pages 86–97. Springer Berlin Heidelberg, 2000. ISBN 978-3-540-67382-8.

[32] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-848-0.

[33] Alisa Kongthon, Choochart Haruechaiyasak, Jaruwat Pailai, and Sarawoot Kongyoung. The Role of Twitter During a Natural Disaster: Case study of 2011 Thai Flood. In *Proceedings of Technology Management for Emerging Technologies*, PICMET, pages 2227–2232. IEEE, 2012.

[34] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A Few Chirps about Twitter. In *Proceedings of the First Workshop on Online Social Networks*, WOSP '08, pages 19–24, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-182-8.

[35] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8.

[36] Hady W. Lauw, Alexandros Ntoulas, and Krishnaram Kenthapadi. Estimating the Quality of Postings in the Real-Time Web. In *Proceedings of the 2010 SSM Conference*, 2010.

[37] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 227–236. ACM, 2011.

[38] Stanley Milgram. The Small World Problem. *Psychology Today*, 2(1):60–67, 1967.

[39] Vincent Miller. New Media, Networking and Phatic Culture. *Convergence: The International Journal of Research into New Media Technologies*, 14(4):387–400, 2008.

[40] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-908-1.

[41] Sushmita Mitra, Sankar K Pal, and Pabitra Mitra. Data Mining in Soft Computing Framework: a Survey. *IEEE transactions on neural networks*, 13(1):3–14, 2002.

[42] Sidharth Muralidharan, Leslie Rasmussen, Daniel Patterson, and Jae-Hwa Shin. Hope for Haiti: An analysis of Facebook and Twitter Usage During the Earthquake Relief Efforts. *Public Relations Review*, 37(2):175 – 177, 2011. ISSN 0363-8111.

[43] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah C. Alhadi. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In *Proceedings of the 2011 ACM Web Science Conference*, WebSci'11, pages 1–7. ACM, 2011.

[44] Daniel M. Oppenheimer. Not so Fast! (and Not so Frugal!): Rethinking the Recognition Heuristic. *Cognition*, 90(1):B1 – B9, 2003. ISSN 0010-0277.

[45] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC'10, Valletta, Malta, 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

[46] Huan-Kai Peng, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. Retweet Modeling Using Conditional Random Fields. In *11th International Conference on Data Mining Workshops*, ICDMW, pages 336 –343. IEEE, 2011.

[47] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. RT to Win! Predicting Message Propagation in Twitter. In *Proceedings of the 2011 International Conference On Weblogs and Social Media*, ICWSM, 2011.

[48] John C. Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical report, Advances in Kernel Methods - Support Vector Learning, 1998.

[49] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowroft. In the Mood for Being Influential on Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third international Conference on Social Computing*, PASSAT, SocialCom, pages 307–314. IEEE, 2011.

[50] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8.

[51] Abraham Silberschatz and Alexander Tuzhilin. On Subjective Measures of Interestingness in Knowledge Discovery. In *KDD*, volume 95, pages 275–281, 1995.

[52] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *2010 IEEE Second International Conference on Social Computing*, SocialCom, pages 177 –184. IEEE, 2010.

[53] Marc Sumner, Eibe Frank, and Mark Hall. Speeding up Logistic Model Tree Induction. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 675–683. Springer, 2005.

[54] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the Right Interestingness Measure for Association Patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 32–41, New York, NY, USA, 2002. ACM. ISBN 1-58113-567-X.

[55] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The Anatomy of the Facebook Social Graph. *CoRR*, abs/1111.4503, 2011.

[56] Ibrahim Uysal and W. B. Croft. User Oriented Tweet Ranking: a Filtering Approach to Microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2261–2264. ACM, 2011. ISBN 978-1-4503-0717-8.

[57] Will Webberley, Stuart Allen, and Roger Whitaker. Retweeting: A Study of Message-Forwarding in Twitter. In *First International Workshop on Mobile and Online Social Networks*, MOSN'11, pages 13 –18. IEEE, 2011.

[58] Will Webberley, Stuart M. Allen, and Roger M. Whitaker. Inferring the Interesting Tweets in Your Network. In *2013 Third International Conference on Cloud and Green Computing*, CGC, pages 575–580. IEEE, 2013.

[59] Christopher Wilson and Alexandra Dunn. Digital Media in the Egyptian Revolution: Descriptive Analysis From the Tahrir Data Sets. *International Journal of Communication*, 5:1248–1272, 2011.

[60] Yunjie Xu. Relevance Judgment in Epistemic and Hedonic Information Searches. *Journal of the American Society for Information Science and Technology*, 58(2): 179–189, 2007.

[61] Zi Yang, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang, and Zhong Su. Understanding Retweeting Behaviors in Social Networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1633–1636, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5.

[62] Aron Yu, C. V. Hu, and Ann Kilzer. KHYrank: Using Retweets and Mentions to Predict Influential Users, 2011.

[63] Reza B. Zadeh, Ashish Goel, Kamesh Munagala, and Aneesh Sharma. On the Precision of Social and Information Networks. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN'13, pages 63–74. ACM, 2013.

[64] Tauhid R. Zaman, Ralf Herbrich, Jurgen V. Gael, and David Stern. Predicting Information Spreading in Twitter. In *Proceedings of the 2010 Workshop on Com-*

putational Social Science and the Wisdom of Crowds, volume 104 of *NIPS*, pages 599–601. Citeseer, 2010.

[65] Dejin Zhao and Mary B. Rosson. How and Why People Twitter: the Role that Micro-Blogging Plays in Informal Communication at Work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, pages 243–252, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-500-0.

[66] Jiang Zhu, Fei Xiong, Dongzhen Piao, Yun Liu, and Ying Zhang. Statistically Modeling the Effectiveness of Disaster Information in Social Media. In *Proceedings of the 2011 Global Humanitarian Technology Conference*, GHTC, pages 431 –436. IEEE, 2011.