The Detection and Inference of the Interestingness of Information in Online Social Networks

Will Webberley

2013

Cardiff University
School of Computer Science & Informatics

Abstract

Abstract text here etc.

Contents

List of Figures

List of Tables

List of Acronyms

GPS Global Positioning System

Glossary

Follower A follower is a type of user. A user, *x*, is a follower of user *y* if user *x* followers user *y*. Other users who follow a particular user will receive all of the user's Tweets and retweets to their home timeline. A user can elect to follow another user.

Friend The inverse of follower. A friend is a user that another user follows.

Path-length The penetration of a Tweet - i.e. the number of times a Tweet is retweeted down one chain. The final retweeter in the chain indicates the number of hops the Tweet has taken from its author.

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.
v. - The act of replicating a Tweet. A user who finds a Tweet interesting may retweet it so that it gets more exposure.

Retweet Group Set of Twitter users responsible for the propagation of a Tweet. Comprises the original author of the Tweet and the users which have retweeted it.

Retweet Volume The number of times a particular Tweet has been retweeted.

Timeline A collection of Tweets in Twitter in reverse-chronological order. A user's timeline consists of that user's Tweets. A user's home timeline consists of the Tweets of each friend of the user.

Tweet *n.* - A piece of information in Twitter; a piece of text, less than 240 characters long, which is written by a user. When sent, the Tweet is sent to the home timelines of

each of the followers of the Tweet's author.

v. - The act of writing and sending a Tweet.

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.

Chapter 1

Introduction

1.1 Twitter and Retweeting

Introduction to Twitter, act of Tweeting and retweeting information. Ideas:

- Twitter is an online social network
- Microblogging - short, quick updates - useful for growing mobile domain
- Used for quick information consumption in spare moments
- Retweeing act of forwarding on of information to those who follow you

1.2 The Problem

Explain about the problem of 'noise' in Twitter:

- Not all information is interesting
- Noisy tweets can dampen experience - becomes harder to find interesting information
- Users cannot find what they do not know exists (similar to Google search bubble)
- etc.

1.3 Contributions

- How is this work novel?
- What benefits does this research provide?
- Does the work solve The Problem?
- etc.

1.4 Thesis Structure

Break down the structure of thesis (i.e. refer to contents page, but also a general overview of the order of sections and what is discussed).

Chapter 2

Background

2.1 Domain Context

Go into more information regarding information propagation in Twitter and about the mechanics of Twitter. Ideas:

- communities and how networks form
- information quality
- information relevance
- information retrieval
- information filtration
- ‘search bubbles’
- propagation
- categorise above into subsections (possibilities: the network and communities, information retrieval and relevance, information retrieval and bubbles, and retweeting as a form of propagation)

2.2 Research Motivation

Link to 'The Problem' in introduction chapter. Talk about wanting to allow people to be exposed to information that they are *likely* to find interesting based on the interestingness of the tweet, but without them having to search for the information or follow the users responsible for sourcing or forwarding the information.

Further work would be done on refining this based on a per-user basis (i.e. that user's particular interests as a relevance metric for the interestingness of the Tweet).

Chapter 3

Related Literature

Either overall lit-review here, individual ones for each chapter, or both (i.e. generall overall and specific ones too).

3.1 Twitter and Retweeting

References for work that helped with the introduction and initial understanding of Twitter. Include papers on:

- Twitter communities and network
- Influential users (‘hubs’ and ‘authorities’)

3.2 Information Quality and Retrieval

References for work regarding quality and relevance of information and retrieval techniques

- Recognition heuristic
- Google search ‘bubble’

- How Twitter handles information quality (Twitter noise, relevance) (evangelists and detractors - Bighona)
- Document retrieval and recommendation systems (Gavalas)

3.3 Tweet Quality and Interestingness

Work on other approaches of defining or measuring Tweet quality and interestingness:

- Approach based on a user's likelihood of retweeting Tweets of a certain type
- Paper outlining the model we used (simulating retweet decisions in a network)

Chapter 4

Understanding The Behaviour of Retweeting in Twitter

Stuff to finish up in this section:

- Explain motivation for research in this particular area
- Use this motivation to explain the purpose for this research as a basis for the work in the next few chapters
- Explain how this chapter is the basis for research into Twitter's social structure in the next chapter
- Normalise terms (retweet-group size / retweet volume) here and in further chapters throughout thesis
- (e.g. we have addressed tweet quality in terms of propagation, can a network have a quality too? what further factors can affect the dissemination of information in social networks?

Social networking and communication now make up a significant part of the Internet, with sites such as Facebook and MySpace attracting millions of users worldwide. Blogging websites, such as Tumblr and Wordpress, have also seen a significant influx of users, allowing people to update and share posts and links on their interests and

every day lives. However, as the need for quick and short updates increases, microblogging sites, such as Twitter, have become more and more popular.

Microblogging is a form of blogging in which posts are limited to a specific character count (140 in the case of Twitter). This limit means that users tend to post updates more frequently [17]. The furthering development of applications in the mobile domain, particularly for iOS, Android and Symbian systems, mean that users update from wherever they are and whenever they like.

Typically, users will tweet of topics that interest them. This may be related to their work, a hobby, or a mixture of multiple areas. 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. Zhao et al. [17], showed that Twitter is also used as a means to contact friends and to get assistance and opinions on topics. Therefore, particular users may belong to different communities of people depending on what kind of posts they want to view.

The strength of Twitter is in its social structure, where users can elect to follow others. Followers of a user receive all of that user's posts in their individual (or 'home') timelines. If a user has set their profile to be public, then their posts also appear on the public timeline, which is accessible to anyone; even those without a Twitter account. As a result, people are likely to follow users who update with interesting posts; whether the follower is a big fan of the friend (a 'friend' is said to be a *followee* of a user) and simply wants to know everything going on in their life, or if the follower is simply interested in the topical area of most of the friend's posts. If someone sees a post that they feel would be interesting to their followers, they can 'retweet' the post, which then relays it as a tweet.

The friends of a particular user effectively become *filters* of information for that user. The user can choose to follow another user, and therefore implicitly indicates the kind of information they want to receive. If you, a Twitter user, want to gain more attention by means of followers, then by posting interesting posts (or, at least, posts that others want to read), you will

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

2. increase the chances that users will decide to retweet your post, thus broadcasting your tweet to a larger audience. People viewing the retweet then may decide to follow you - the original tweeter.

It is, of course, possible that more than one user may decide to retweet one of your posts. A follower of a retweeter may decide to retweet the retweet, creating a chain. Naturally, the larger your effective audience (both directly and through retweets), the greater chance, again, you have of being retweeted. [13] showed this by demonstrating how the retweet rate increases with the number of followers of the original tweeter. This is also related to the ideas of user influence mentioned in [4]; that more followers implies more influence. The retweeting process ultimately results in a tree system, with the root being the original tweeter, and other nodes being retweeters at different levels, similar to the idea of URL cascades discussed in [7].
The motivation for this study has stemmed largely from the interesting decentralised nature of retweets, spurred on by similar literature such as information cascades in [7]. The notion of retweeting is also very similar to content forwarding in opportunistic networks, particularly ideas discussed in [1], and studying the menchanisms of retweeting from a human-centric perspective may provide insight into protocols for autonomous content dissemination. We hope that understanding retweet properties may help us in forwarding relevant information to a user from without that user's social neighbourhood.

4.1 Retweet Trees

Example of a retweet tree.
Source users are roots, final retweeters are leaves (although subject to change as time goes by).

Links between nodes indicate retweeting from a user to another user (but not which users follow which).

Define path-length and retweet groups.

For this paper, the term *path-length* is to be defined as the number of times a particular tweet is retweeted down a chain. For example, the status;

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

has a path length of two. All users involved in this retweet chain have used the old method of retweeting, which is still very commonly used and has effectively provided most of the results for this paper.

4.2 Information Retrieval

People don't use Twitter to obtain any specific information (they don't know what people are necessarily going to say)

If all friends are technology based, then you can expect technological tweets, but generally this isn't the case.

4.2.1 Epistemic Search

4.2.2 Hedonic Search

4.2.3 Affective Stimulation

4.2.4 Search 'bubble'

4.3 Background to This Area

To date, research has largely focused on the actual structure of Twitter in terms of the friend-follower graph [**?**], user communities [10] and also the psychology behind Twitter users [17] [4]. A lot of these studies link well with retweeting, as will be explained later, and researching them has helped to explain and facilitate retweet studying at a wider angle.

Arvola et al. [2] introduced a retrieval system for online documents which returns particular passages in documents which are of a pre-defined interest for the user (generally through the use of search terms). In Twitter, a user can specify what they are interested in receiving by following a set of users, with their neighbours effectively becoming the search term. Twitter then enables the user to receive more relevant and interesting information based on these filters. However Twitter doesn't alter the reading order of passages (or tweets) as in the aforementioned retrieval system, the home timeline only shows the tweets coming from people the user has chosen to follow. [16] discusses user's relevance judgement based on hedonic and epistemic search, particularly linking to the idea of 'affective stimulation', occurring when the user becomes affected by some information (e.g. emotionally or for entertainment). This is somewhat also applicable to following people in Twitter as users don't generally use Twitter to receive any particular form of information, thus they haven't *searched* for any specific topic. Therefore, any information they do receive, and which they find interesting, be-

comes affective stimulation. If someone follows many users, then he/she is receiving too much information to effectively filter out the interesting information. If following a few users, he/she can't be receiving *all* the information that they might want to.

However, despite the fact that this system is thus quite sound (by largely displaying information the user *chose* to receive), it may not be particularly complete since it is likely that there are many other users and topics the user might still be interested in which is not returned by this Twitter information retrieval system. [2] also discusses the notion of expected reading effort and its importance in information retrieval systems. The character-precision recall metric was introduced as a way of demonstrating the tolerance-to-irrelevance ratio. The general mechanism for this is to do with users reading a document passage; the point the user reaches this ratio is when the user stops reading the particular passage and moves to the next whole document (since they assume the rest of the document is also irrelevant to them). Therefore, the more effective the information retrieval system, the lower the chance that this ratio is reached by the user. Associating this with Twitter, we can say that a user viewing tweets from a user they are following may get to the point where he or she reaches this ratio (i.e. is bored or finding the tweets irrelevant) and decides to unfollow the friend. Similar to above, the more effective the user is when selecting people to follow (friends) in the hope of receiving interesting information, the lower the chance that the user will remove these friends.

As discussed in the introduction, a user is more likely to follow other people who post tweets of interest to the user. As more and more people follow a particular user and start tweeting a lot about a particular topic, the users may start to follow each other, producing a growing 'swarm' of interest around a certain topic (or a particular user). As more users start associating with this user or topic, its audience becomes more widespread, attracting further users to it. As this continues over time, a community of users can be formed. This leads onto the notion of Twitter communities discussed and experimented on by [10], who define communities to be groups of users with dense follower links between them. Their studies continue into looking at the development

of communities over time and also defines how they can be formed from different types of people - such as those who tweet lots and have many followers, and those who post little and have few followers. Studies in the behaviour of different types of users in Twitter is done more thoroughly in [**?**], where 'broadcasters' - users with many followers and few friends - and 'miscreants' - users with few followers but many friends - are discussed.

Applying the community idea to retweeting, it seems likely that a user in a community is going to retweet posts from other users in the community (due to the shared interests). This is useful when the user has followers outside of the community, since they are then less likely to be following the source user. However, if the retweeter has many followers inside the community, then they are similarly likely to be following the source tweeter and so would be exposed to the original tweet anyway, since there are more interlinking edges between users in the graph. If many users within the community end up retweeting a post, then other users in the community may end up be exposed to the same tweet retweeted several times. This paper does not study this problem, but more analyses the behaviour of retweeting in general, paying particular attention to any trends or patterns arising.

4.4 Twitter Analysis

Discuss wanting to know more about Twitter and the way information is propagated through this.

Want to gain a base knowledge of some of the propagation patterns in terms of the tweets and the networks the tweets travel through.

This will form basis of research further on.

Stuff interested in:

- retweet groups (define)

- path-length (define)
- friend/follower graph
- timings, etc.

4.5 Experimental Methods

4.5.1 Data Collection Methodology

The experiments conducted involve the examination of tweets selected randomly from the Twitter public timeline[1]. The system interfaces with the Twitter REST API and periodically retrieves the top 20 tweets from the timeline, filtering out the posts that are retweets. They are distinguishable as they start with the characters 'RT' followed by a username. These retweets are then stored and the original tweet text is extracted. Sometimes retweets are made with additional text as comments, which made it harder to retrieve the actual original text. In these cases, additional queries are made to try and source the original text, but, failing that, the retweet is discarded.

The original text is then used to query all available tweets for the original tweet and any other retweets. The original tweet, along with its set of retweets is referred to as the *retweet group*. As a result, each retweet group contains precisely one original tweet and at least one associated retweet, thus the smallest possible retweet group is of size two. The *final retweeter* is the most recent person to retweet the post. Results are generated by extracting data from these retweet groups, which are stored with metadata such as time, source, language, and so on.

It should be noted that there are two ways in which tweets can be retweeted. The traditional (or *old*) manner involved users copying the existing tweet (or retweet) from the timeline before tweeting that with the characters 'RT' and the previous retweeters username at the start. This ensures that the appropriate users are cited as being the

[1] Viewable from *http://www.twitter.com/* when not logged in.

author of the tweet. The *new* method enables users to retweet a post simply by clicking the 'Retweet' button that is available next to the post on the timeline both on the main Twitter website and on various Twitter clients. Despite the fact that the website and clients have alternate ways of displaying these types of retweets, the search interfaces and API still see them as starting with 'RT' and the previous poster's username. This, fortunately, means that both types of retweets can be collected in the same way. Retweeting in the latter method, however, is unavailable if the post that is to be retweeted belongs to a user who has a non-public profile.

Despite this, the new method doesn't have direct support for retweet chains (explained below) to be formed. No matter who retweeted the post previously (whether they were the original poster, or not), the new retweeting method simply treats it as a direct retweet from the original poster (as long as the previous retweets were also made in this fashion). The original tweet (and other retweets) can still be found using the search method outlined above, however, so this doesn't significantly affect results. Limitations with both the Twitter API and the search interfaces did mean that not all retweets could have successful queries made against them.

As briefly described above, these experiments use the notion that retweets themselves can be retweeted; analogous to routers forwarding packets in a multi-hop networking system.

4.5.2 Data

There are three sets of experiments detailed in the following section and all of the data was collected using the search method outlined above. The data consists of around 4400 retweet groups (representing a total of about 26,000 tweets and retweets) collected between 26th January & 24th May 2011. It is acknowledged that this is a relatively small result set and so we emphasise that these are *preliminary* experiments to provide insight into the decentralised side of Twitter and to act as a start point for future work

in this area.

4.6 Empirical Results

4.6.1 Length of Retweet Chains

The longest path-length encountered from the dataset of retweet groups was 9, meaning that the tweet has been forwarded nine times from the original user; making the total number of involved users ten. Despite this, in the dataset, the average path-length was around 1.8, with the vast majority of tweets being retweeted between one and two times, as shown by the distribution in Figure 4.1. The *maximum path-length* of a retweet group refers to the longest path-length present in the group.

It should be noted that the very high proportion of single-chain path-lengths could be

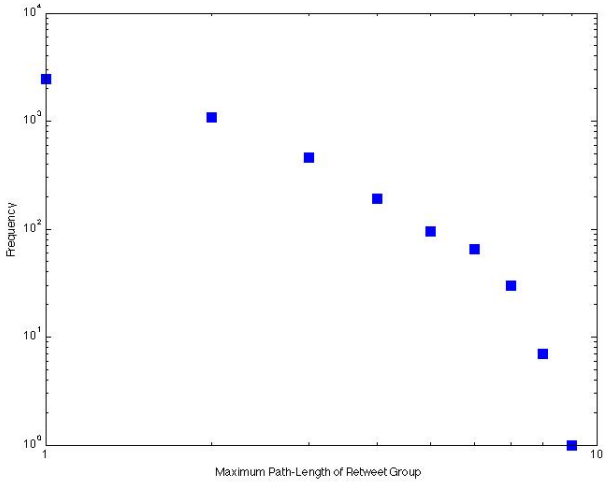

Figure 4.1: ***Log/log distribution of maximum path-lengths from retweet groups.***

a result of the new method of retweeting introduced in the previous section, since, as mentioned, if all of the users in a retweet chain were to retweet in this way then Twitter would still only treat the path-length as one; between the original tweeter and each of

the intermediate retweeters. Despite this, the result fits the rest of the dataset well.
This data can also be used to have a look at the friend-follower graph associated with retweets. In cases where the group's maximum path-length is equal to one (i.e. simply the case where a user has retweeted another user's tweet once), the retweeter follows the original tweeter about 90% of the time. This implies that in 10% of cases, a retweeting user has retweeted a tweet found outside of their home timeline, i.e. on the public timeline, or whilst browsing through other users. It could also be as a result of the retweeter not including the tweet citation ('@username') for reasons such as to save space, which would be particularly likely in cases of longer tweets. This implies that the more followers a user has, the greater chance that user has of having his or her tweets seen and of being retweeted. Since 90% of retweets occur if the retweeting user follows the original tweeter, then this is directly demonstrated. However, this does not take into account whether the tweet was passed down a chain of retweets retweeted with the second (single click) method. In these cases the path-length would be represented with a length of one even if they were actually longer. The introduction of the second method of retweeting is, therefore, not helpful with the analysis of retweets.
In addition to this, in cases where the maximum path-length is greater than one (i.e. at least one user retweeting in a chain between original tweeter and final retweeter), the final retweeter follows the original tweeter in about 40% of cases. From Figure 4.3 we can see that retweet groups with a larger maximum path-length tend to be larger themselves. This means that the tweet has travelled further both around the original tweeter, but also potentially to other communities. Users from other communities will be less likely to follow this original tweeter, explaining this drop in likelihood. Results have also hinted that the probability of users citing the original tweeter decreases as the path-length of the retweet increases, which also goes some way to explain this point.

4.6.2 Size of Retweet Groups

This section focuses on properties of retweet group size, and also how this can be related to path-length. The average retweet group size found from the data was of size 6, with the largest and smallest retweet groups being of size 284 and 2 respectively.

The distribution of retweet group sizes can be seen in Figure 4.2 and shows that this data follows a power-law distribution (with a relatively large p-value of 0.871). The $Pr(X \geq x)$ represents the complementary cumulative distribution function, where the randomly generated X is greater than or equal to x. These values were calculated using the techniques and code provided by [6].

In these retweet groups, and particularly the larger ones, retweets are likely to be down several chains. Combining this information with that of the average path-length of retweets, it becomes more clear how the retweet tree structure is created, with nodes being the retweeting users (the root node being the original tweeter), and edges between nodes being the path of retweets. Different retweet chains would be illustrated by the combinations of journeys down different branches. The longest path-length would be represented by the depth of the longest tree branch, and the total number of retweets would be represented by the number of edges. The trees can be said to represent the retweet groups, though, in practice, many members of the group will be separate from the main tree due to missing citations.

[11] discusses retweet trees in more detail, in particular, the different shapes of the trees representing retweet groups from the Air France incident. Their results, for these trees, also (see section 4.6.1) indicate a large number of groups with maximum path-lengths of one and two.

Figure 4.3 shows how the total number of retweets varies with the longest path-length of the retweet group. The trend mostly correlates with what might be expected; that the maximum path-length of a group increases with the overall size of the group. These illustrations do not show which users are followers of others, but do show how some

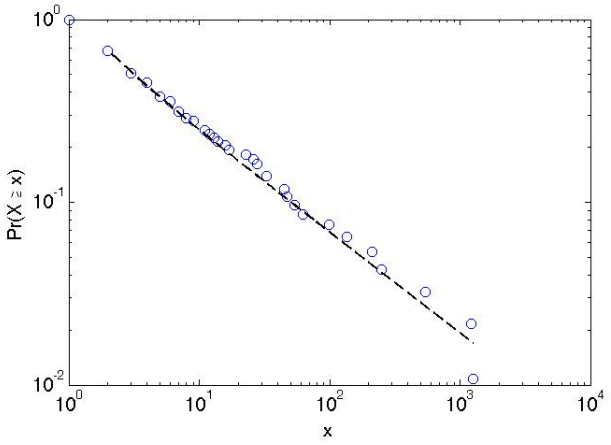

Figure 4.2: ***Maximum likelihood power-law fit for the cumulative distribution of retweet group sizes..***

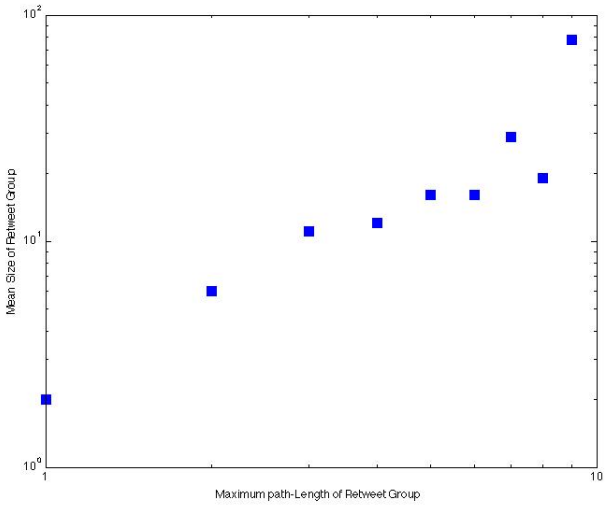

Figure 4.3: ***Log/log relationship between the maximum path-length and size of a retweet group..***

tweets are retweeted significantly more than others. Users who have many more followers are said to be more *influential*, say Cha M. et al. [4], who also discuss the idea of 'indegree', and that those users who have far more followers than friends are likely to be far more influential. Their chance of having tweets retweeted is therefore increased.

The (immediate) audience size of a retweet group refers to the number of users the tweet has been exposed to either directly or through retweets. *Immediate* was used in this sense since users who have made their profile public can have their tweets viewed by users who aren't followers of the former. This audience size can therefore be calculated simply by summing up all the followers of each of the users in the retweet group.

These audience measures take into account that some users may be exposed to the

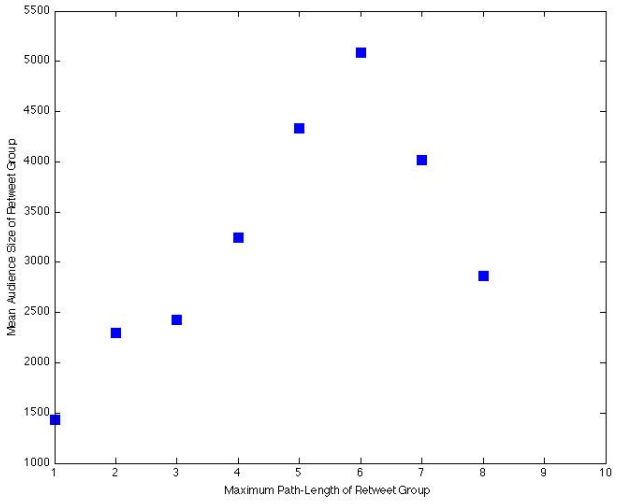

Figure 4.4: ***Relationship between a retweet group's audience size and its longest path-length..***

same tweet more than once. This happens when users involved in a retweet tree have some followers in common, and so is likely to be prevalent in more closely-knit communities. As a result, the audience sizes shown only represent *distinct* users exposed to the tweet.

Collection of the audience size from the data started slightly later than the main dataset and is only available for 2860 of the total 4400 groups. The longest maximum path-length of this subset is 8.

Figure 4.4 shows how the audience size of retweet groups varies with their maximum path-length. The peak at path-length 5 indicates that the groups with a mid-range maximum path-length tend to have a larger audience size, and we believe this is to do with

the amount of audience overlap in different retweet groups.

[11] discusses how retweeting is related to the audience size of a tweet and how the power of the retweet phenomenon can greatly affect the spread of information, even if the original tweeter 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 tweeter. This can also be seen in our results; that no matter the maximum path-length of a tweet, the number of users reached is relatively high.

The overhead of a retweet group represents the number of users who are exposed to a tweet more than once and is present in 71% of retweet groups. The *proportionate* overhead is the ratio of overhead to audience size and the mean of this was found to increase with the group's maximum path-length. This is a possible explanation for the peak in the data: that eventually the overhead of non-distinct users has increased to the extent that it reduces the audience size more significantly. The same graph representing the effective audience size (calculated with the addition of non-distinct users) represents, mostly, a continuous positive correlation.

Three of the largest five overheads collected were from retweet groups with a maximum path-length of 1, the largest with an overhead of 6.5 times greater than the distinct audience itself (the overhead was larger than the actual audience size in around 3% of retweet groups). This shows that, to an extent, there can be significantly more overlap in more closely-knit communities; those retweet trees which are wide and shallow. The chance of getting no overhead increases in smaller retweet groups.

4.6.3 Retweet Follower Pattern

This experiment focuses on the pattern of followers in the retweeting hierarchy.

The first result shown from the experiment is that the final retweeter follows the previous retweeter in the chain in 67% of cases. It initially seems strange that this should be 20% lower than when following a user in retweet chains of length one. This sug-

gests that users involved in shorter-chain path-length retweets are members of more tightly-knit communities. Retweets with longer path-lengths have, by nature, travelled further and so would be the type of retweet to travel between communities, reducing the chance of the involved users following each other.

The interesting part of this, however, is the number of followers of the previous retweeter in different cases. In the 33% of cases where the final retweeter doesn't follow the previous retweeter, the latter has, on average, around 600 followers. When the final retweeter *does* follow the previous retweeter, however, the previous retweeter's average number of followers is 940. This is quite a substantial difference and certainly highlights the fact that by having more followers you are more likely to have more influence in terms of whether you get retweeted, or not.

This is accentuated further when looking at the original tweeter. The likelihood of a retweeter following the original tweeter in cases in which the path-length is of more than one has already been found to be around 40%, but the average number of followers of the original tweeter increases by a factor of around four (580 to 2000) when also followed by the final retweeter. Results showed that the original tweeter had a consistently higher number of followers when followed by the final retweeter than when not at all path-lengths. This demonstrates that having an increased number of followers is correlated with the chance of a user being retweeted. In this case, having four times the followers increases the correlation dramatically (40% to 90%). The number of followers of a user can therefore be directly related to the ideas of influence discussed in [4] and also of 'advertising' themselves.

It was found, however, that the number of followers of the original tweeter diminishes as the path-length of the tweet increases (Figure 4.5), signifying that tweets travel further when the original tweeter has fewer followers. Because the retweet groups were collected in such a way so that groups containing longer path-length retweets also contained many shorter-chain retweets, retweet groups containing path-lengths of 5 (or more) are also likely to contain many retweets (if not more) with path-lengths of one or two (see the distribution in Figure 4.1). It can therefore be argued that there are

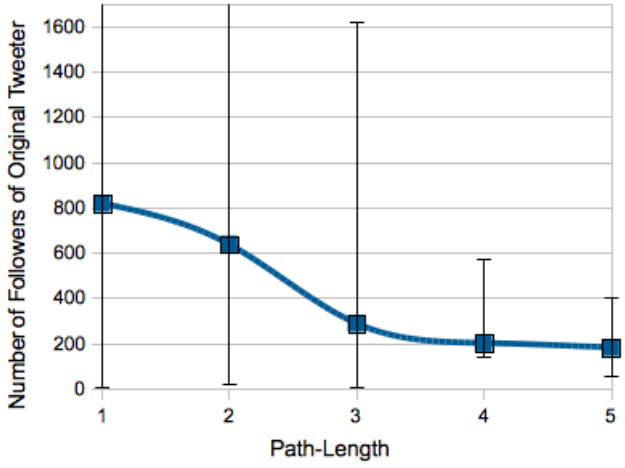

Figure 4.5: ***Relationship between number of followers (and respective distribution) of the original tweeter as the path-length increases..***

more users involved in shorter-chain retweets than in ones with longer path-lengths. It is then more likely for these users to have more followers than others in the retweet group. Another explanation could be that users are actually aware of their local network and realise that retweeting may cause a lot of audience overlap (particularly in the case of large communities). A user may have seen a post retweeted a few times on their home timeline and thus decide not to also retweet.

One last interesting point to make regarding the notion of retweet chains is looking at the how the pattern of following previous retweeters develops as the path-length increases. It has already been discussed above how the chance of following the previous retweeter in the chain is about 67%, but, in cases where the path-length of a tweet is greater than two (i.e. at least two intermediate retweeters between final retweeter and original tweeter), the chance of the final retweeter following the next retweeter along preceding the previous retweeter is around 45%. This suggests that retweeting is more widespread and not so much just circulated around communities. These preliminary results demonstrate that the chance of the final retweeter following previous retweeters - up to and including the original tweeter - diminishes along the chain or as the tree is ascended (Figure 4.6).

Because of this, it's sensible to assume that the tweets in the dataset are forwarded

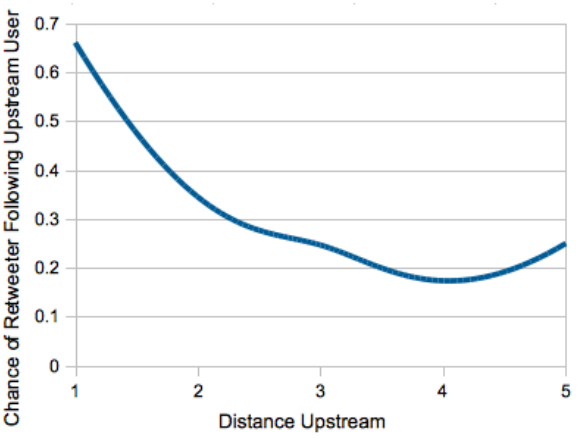

Figure 4.6: ***Proportion of final retweeters following upstream users at varying distances along the chain..***

through less-connected users, and perhaps forwarded from community to community by those users belonging to several groups. Otherwise, if the retweets were circulated more around closely-knit communities, the likelihood of the final retweeter following the previous tweeters would be both greater and more evenly spread - i.e. the chance of following the previous retweeter would be roughly equal to the chance of following the other tweeters in the chain.

In addition, of the 67% of final retweeters who *are* following the previous retweeter, about 19% of them also follow the next previous retweeter (i.e. the retweeter at path-length - 2). In this case, the next previous retweeter has, on average, 3000 followers. In the 81% of these users *not* following the next previous retweeter, then the latter has an average of 525 followers. This is an accentuated result of the one previously, but this time boasts an increase of a factor of 6.

Of the 33% of users who *don't* follow the previous retweeter, about 30% follow the next previous retweeter. Both of these sets of statistics also go towards the idea of the diminishing chance of following the users as the tree is ascended.

From this dataset, it was also possible to work out how often retweeters cited the original tweeter of a post. In retweets, users are typically cited by, as we have seen, having their name along with an 'RT' at the start of the post. This data was collected by seeing

if the original poster's username was mentioned *anywhere* in each retweet. The chance of this occurring was found to be around 68% and did not vary with any pattern with path-length.

4.6.4 Retweet Time Delay

The final experiments in this section focus on the time delay between the final retweeter and original tweeter. This is an interesting area since it enables researchers to see how fast messages propagate through the Twittersphere. From this information, and by using the retweeter patterns demonstrated above, it would be possible to work out how far and how quickly information can be passed around.

Figure 4.7 shows the average time delay between the first and final retweet with increasing maximum path-length of the retweet group.

The results indicate that, mostly, as the group's maximum path-length increases, then

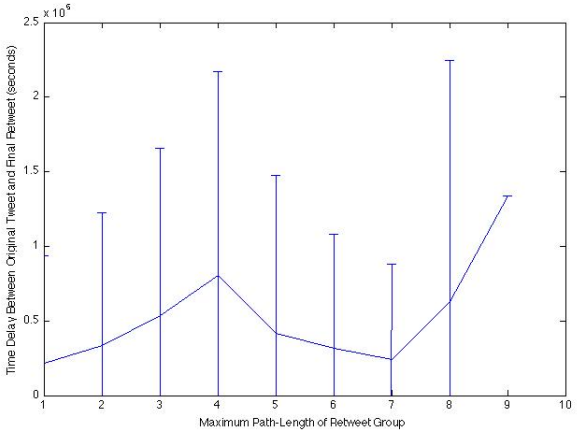

Figure 4.7: ***Average time (in seconds) between first post and final retweet of a retweet group varying with the group's maximum path-length..***

so does the elapsed time between the original post and the final retweet. This is probably as was expected, since this shows that it takes longer for a retweet to travel further. The data is not consistent, however, especially results for a path-length of five and above. The first four results suggest a uniform incline roughly proportional to $v = \frac{s}{t}$,

where the distance, s, is the hypothetical distance given by the path-lengths, showing that the *speed* of propagation remains mostly constant.
There is not enough of a trend in the data to make any deductions regarding propagation speed, however. There are two main conflicting arguments regarding this result: the first is, as mentioned, that the further a tweet travels the longer time it travels for. The second is to do with tweet popularity: the more popular a tweet is, the more quickly it will be retweeted. In the latter case, it is possible that longer trees grow fully before shorter ones, implying exponential growth. Generally, though, it seems that the maximum path-length of a retweet group does not massively affect the tweet's propagation speed.

4.7 Summary

The experimental results have certainly highlighted the ideas of communities and that of message cascading similar to that demonstrated in [10] and [7] respectively.
As has been seen, in section 4.6.4, the retweet tree seems to grow in a variety of ways. One argument is mostly expected; that as the 'distance' the tweet travels increases, then so does the time taken for it to reach its end. The other argument is linked to the idea of tweet advertising, discussed in previous sections, and namely the notion of tweet popularity.
The previous experiments showed how the number of followers of a user directly influences their chance of being retweeted. It can therefore be seen that advertising can also be linked to the level of influence a user has. The results help illustrate the multi-dimensional properties of the retweet tree, and how these factors relate to its growth and its associated friend-follower graph.

This paper has demonstrated relatively simple results in an attempt to realise some of the behavioural patterns of retweets, both linking to the psychologically in terms of the users, but also the physical properties of retweets. The results are able to repres-

ent a basis for potential further work and research into the various aspects of Twitter, moreover, perhaps, topical categorisation and the dynamicity of the friend-follower graphs.

Chapter 5

Analysis of Twitter's Social Structure

Stuff to add to this section:

- Change tweet features for each simulation and make comparison on these differences
- Observe differences in patterns when network generation parameters are altered
- Link up section to previous section (i.e. how did the previous research help and how does this build on that work?)
- Explain how this section becomes the basis for work in 'main chapter 3'.
- (e.g. Issues with current method (too long, requires network, inaccurate due to having to choose users with fewer followers), so need a quicker, more accessible and online approach).
- Explain the Mechanical Turk questions in more detail, with examples.
- Discuss about the machine learning approach used (logistic regression and how it works)
- Link 'retweet volume' to 'retweet group size'

Twitter is often seen as one of the biggest sources of new and live information on the Internet, with millions of people producing and absorbing information daily. Users

receive tweets onto their timelines from the users that they follow. Thus, a user has some control over the *type* of information they receive by choosing which other users to follow. A particular user, therefore, may not be aware of information that exists outside of their local network, since he or she is not directly exposed to the information produced by non-followees.

Retweeting allows users to forward information they receive onto their own followers. As a result, followers of these users, who might not normally be exposed to this information, now have a chance to access it. A user who decides to retweet a tweet can be said to consider that tweet to be *interesting* (at least, to their followers), since that user has taken the time to read the tweet, decided whether or not to share it, and then to actually retweet it [14].

Several factors can affect a user's decision to retweet, such as whether the tweet contains an (interesting) URL, whether the tweet mentions another user, whether a user even has a chance to see the tweet, the influence of the author, and so on. These factors account for a user's individual retweet *decision* on a particular tweet, and the combination of several users' retweet decisions dictate how far the tweet will propagate. However, it is our belief that the social network structure also has an affect on how far tweets can travel.

The social structure of Twitter is built up by users electing to follow other users. When a user follows a user, a directional link is forged between them, and any tweets generated, or forwarded, by the followed user are passed down the link. It's clear to see that, as more links are made between users, many more avenues are generated for message propagation throughout the social structure. Users with a high in- and out-degree can become an information highway, but users with a low out-degree are a bottleneck of information.

In this chapter, we demonstrate how different network types support different propagation characteristics through the use of a model simulating each network type. Using the model, we make predictions on the retweet outcomes on several network types, and compare these to the characteristics of the propagation in real Twitter networks. We

finish by discussing how the *interestingness* of a tweet may be inferred from simulating the network in this way.

5.1 'Twitter is a Memepool'

Introduction to memetics (Richard Dawkins intro maybe):

- Gene is a physical entity containing information and instructions. It is a unit of genetic inheritance (i.e. offspring typically have a mashup of the genes of the parens)
- The result of the data held by a gene (the genome) means that organisms with certain genes are able to reproduce and survive more than other organisms containing different genes.
- Thus the gene is able to replicate under certain gene- and environmental-centric conditions.
- A meme is similar to gene but is non-physical. They are a unit of cultural inheritance (an idea, phrase, behaviour, etc.).
- Like genes, memes are able to survive better when their features (*menome*) are suited to the meme's environment. In such environments, the meme is able to be shared and replicated more efficiently and frequently.
- A Tweet, again, is similar to both. A Tweet itself has many features (the text of the Tweet, the time of its origin, its length, etc.) and their environment, the Twitter social structure, has features (namely the users that belong to it and the way they are connected) which may facilitate the replication (i.e. Retweet) of the Tweet.

- A Tweet existing in different social structures will have different Retweet patterns, which is what we want to show in this chapter.

5.2 Background to This Area

[14] discusses an idea similar to that of predicting the interestingness of a tweet, but focuses mainly on predicting the users most likely to find the tweet interesting enough to retweet. Similarly, [9] looks at the same issue, but at the other way around by predicting the *type* of tweets that are likely to be retweeted many times. Discussions on the retweet decision in relation to a user's recognition of its features take place in [5] as well as conclusions about the effect of features such as the number of followers of a user or any pre-existing metadata on the interestingness of the tweet.
The notion of time decay and how this is associated with a user's retweet decision is discussed in addition to a retweet probability prediction in [18] (and [12]), which is the basis of the model we use in this paper.

5.3 Overview

In the next sections, we introduce and briefly explain the regression model we use for simulations and predictions. We then go on to analyse the differences in the propagation characteristics between three different network types before comparing the results of simulations on these networks with data collected from the Twitter social graph. We finish by introducing a methodology for predicting the interestingness of a particular tweet to a particular user and how this might be improved.
Ideally, there are two things we'd like to see from the first set of experiments; firstly, that changing the network type and properties does, indeed, affect the propagation behaviour, and, secondly, that at least some of the results from the experimentation correspond to Twitter's own retweet behaviour so that a fair comparison and justifica-

tions of our results can be made.
With regard to our prediction work, we'd like to be able to make relatively decent predictions on which tweets are of interest, and which are not, based on the simulation research in the next section.

5.4 Model

As mentioned above, [18] introduced and discussed their prediction model, which was shown to perform well when predicting the retweet decision of a user. Their model trains a logistic regression using a set of user-, tweet- and context- features in order to classify an experimental tweet and output a retweet probability based on a set of features.

5.4.1 Algorithm

Include image indicating the flow of events that take place for the model
Pseudocode for the algorithm

5.4.2 Features

We use this model as part of a simulator in order to obtain an approximate retweet volume of each tweet when the retweet probabilities of each user receiving the tweet are combined. [18] used a set of around 50 features, though they mention how some features have more weight than others. In order to simplify the simulation significantly and to make data collection more tractable, we decided to use only the following four major features;

- *follows* - Whether or not the user exposed to the tweet is a follower of the author;

- *followed* - Whether or not the user exposed to the tweet is followed by the author;
- *mentioned* - Whether or not the user exposed to the tweet is mentioned in the tweet;
- *URL* - Whether or not the tweet contains a URL;

Where *author* is defined as the user who originally tweeted the tweet.
In essence, the model requires a network of users and one tweet to start the simulation. It starts by initialising a user set, *U*, to contain one source user from the network, U_s. This source user then transmits a tweet which is then received by its followers. U_s is then removed from U. For each follower, its tweet and user features are classified to produce a retweet probability. If this is greater than a random number, R, then the user retweets the tweet to its followers and the process repeats. Any user that retweets the tweet is removed from U and added to the retweet decision set RT. The followers of the retweeter are added to U for the next iteration, in which each user in U has received the tweet onto their timeline. The retweet volume of this tweet in this network is then the cardinality of the set RT.
A tweet in a user's timeline will slip further down in the timeline as time goes by. This happens whether the tweet is interesting or not and whether or not the user has even seen the tweet. Users having a quick browse through Twitter may not have time to scroll down to find these interesting tweets (and will not know they exist) and thus tweets left in the timeline have their retweet chance decay over time.
We emulate this phenomenon in our simulations by removing the tweet from a user's timeline (by removing the user from the set U) if the user has not retweeted the tweet within a timestep threshold, which can be varied to alter the volatility of the retweet.

5.4.3 Training the Model

To collect the training data, we crawled Twitter, using the REST API, from March-June 2012 to collect a set of around 12,000 tweets and retweets from the Twitter public

timeline. In addition, we made further calls to collect the information required for the features (namely the *following* and *followed* features). The above features were extracted in each case and the regression was trained.

5.5 Network Analyses

In this section, we look at three different network structures and discuss the differences in the propagation patterns produced by each. Note that each graph we assess is *directed*. The same set of tweets are used for each simulation on each of the networks.

5.5.1 Path Network

A path network is the most simple of the three, and is also the least life-like when compared to Twitter's own social graph. In this network, the output retweet volume is, by definition, equal to the penetration (i.e. *depth* of propagation) of each tweet.
In this case, a directional path network consists of a network of users, N, of size n, in which each user N_i is followed by user N_{i+1} for $1 \leq i < n - 1$. As a result, all users in the network, except the user N_n, have precisely one follower.
Since each internal user only has one follower, the likelihood of a retweet occurring at each timestep is somewhat reduced, it is expected that the retweet volume will tail off more soon than in other network types. Propagation is also hindered by the fact that each retweet can reach an audience with a maximum size of 1 at each stage, thus relying on that single user making the retweet. Figure 5.1 shows the frequency distribution for a path network when simulated with the logistic regression over a series of tweets. The graph shows a very large proportion of single retweets, which reduces logarithmically with larger volumes.
The likelihood of a user getting the chance to retweet, and also deciding to retweet, becomes the product of a probability function the further the tweet travels, where user

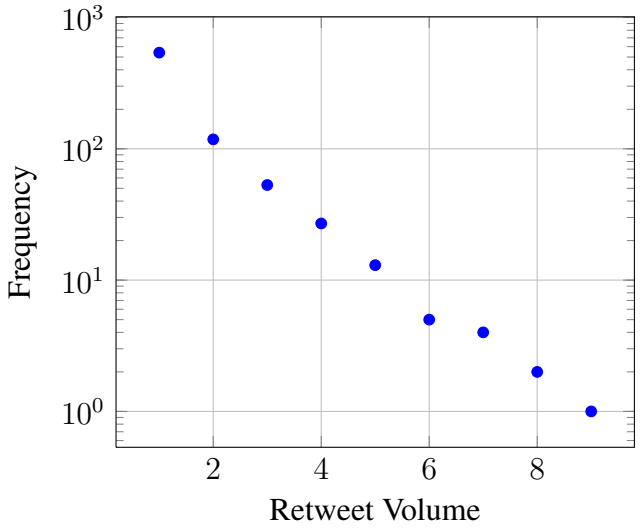

Figure 5.1: Retweet volume frequency distribution from path network simulation.

N_i requires all users N_0 to N_{i-1} to pass on the message before it even gets a chance to make the retweet decision.

As a result, if the retweet decision chance of each user is more or less equal, the chance of user N_2 retweeting the tweet is of an order of magnitude less than that of user N_1 deciding to retweet. The graph shows half life-style behaviour; owing to the fact that each retweet is exponentially less likely to occur than the previous retweet. Note that the graph is plotted on a log-linear scale.

5.5.2 Random Network

Random networks are more similar to Twitter's own social structure than path networks, but are a much more basic and uniform version and do not consider more influential users or the development of Twitter communities.

A random network is defined as a network of users, N, of size n in which a user N_x has probability p of following user N_y. Thus; as the probability p is increased, the likelihood of a user following other users in N increases, causing the overall network edge density to increase. Generally, the average number of followers and followees of

a user is proportional to $p \times n$. Thus the parameters for constructing such a graph are the network size, n, and the attachment probability, p. The simulation results for the random network indicates a higher distribution of mid-range retweet volumes.

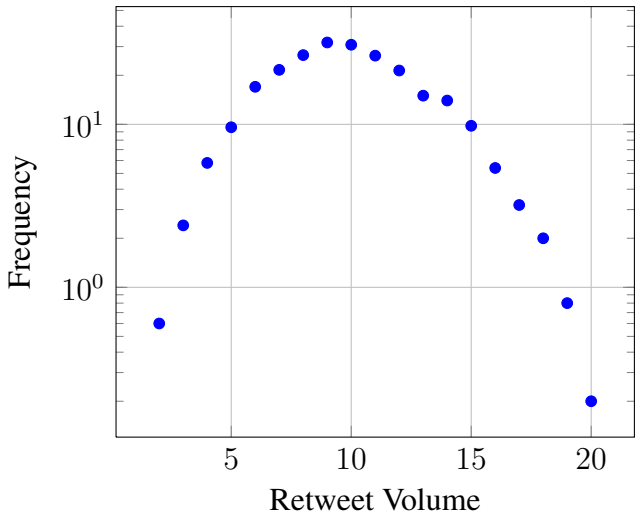

Figure 5.2: Retweet volume frequency distribution from random network simulation.

5.5.3 Scale-Free Network

Include more mathematical analysis of scale-free networks (in general) - i.e., in what way are they logarithmic?

A scale-free network is a network of users, N, of size n and is generated in such a way so that the resultant distribution of degree follows a power-law. *In*-degree signifies the number of inward edges to a node (i.e. the number of followers of a user), whereas *out*-degree is the number of outward edges (i.e. the number of users that user follows). Scale-free networks have been the subject of a fair amount of research, and are explained more thoroughly in [8]. In our implementation we use NetworkX[1], a Python networking package, to generate directed scale-free networks through a preferential-attachment algorithm based on the network size and edge density as parameters.

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

Figure 5.3 shows the frequency distribution of retweet volumes. Since the data is plotted on logarithmic scales, we see a logarithmic trend very similar to our results in [15].

5.5.4 Comparison to Real Twitter Data

In our previous work, [15], we captured and analysed data which contained results on the distribution of retweet group sizes. In that paper, a retweet group was defined to be a set of tweets containing one tweet and then all the retweets of that tweet. Thus the retweet volume looked at in this section is effectively the cardinality of the retweet group (minus one). Since the results in the above experiments also look at the frequency distribution of retweet volumes, then we should be able to draw some comparisons.

We compared the data produced by the different types of network to this previous data, and found that the scale-free network produced a distribution similar to that from the real Twitter data.

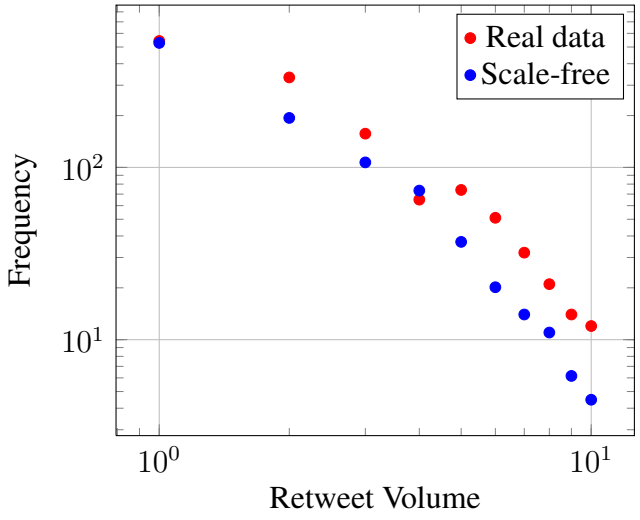

Figure 5.3: Comparing the retweet volumes distribution from scale-free graph simulation to data from Twitter's graph.

5.5.5 Structure Comparison

Each network structure has been demonstrated to show different propagation characteristics. This has shown that, in addition to a user's own retweet decision, the actual spread of a tweet depends somewhat on how the author's local network is constructed. With lots of edges in the graph, there are many more paths down which propagation can occur, increasing the number of times a retweet decision is made, and therefore an increase in the overall number of retweets occurring. The retweet decision facilitated by the model, therefore, combined with a user network give an overall *retweetability* of a tweet that will vary depending on the network it's being propagated in, the source user, and the intermediary retweeters.

The path network, as designed to be the extreme case, has shown to allow poor propagation. Importantly, whilst the network parameters and retweet decision had to be globally increased to obtain any sensible data from this simulation, the trend still shows how propagation down a single chain isn't hugely effective.

The random network facilitated many more retweets due to the fact that users had a very similar in- and out-degree in all cases across the network. This means that each user is able to receive a lot of information, and is also able to pass on (whether an author or a retweeter) information to lots of users simultaneously in the graph.

Despite random networks supporting large retweet throughput (i.e. high *recall* of information), the disadvantage is that the interest *precision* is much lower. This is because this type of network relies on users following a large number of other users, thus meaning that they would receive more 'noise' (i.e. uninteresting tweets) than if they were more limited and selective. Although tweets that are retweeted are usually of a higher *quality*, not all retweeted tweets will be interesting to all users.

Finally, the scale-free network, whilst not having the highest throughput of information, does have trends most similar to the data on retweet distributions collected from Twitter's social graph. This is due to its ability to emulate more influential users and areas of dense communities (as discussed in [10]). These networks have the poten-

tial to allow for large numbers of retweets, especially if they are sourced from one of these more dense areas, but typically die off more quickly as the tweets are retweeted through less influential users.

5.6 Other Findings

5.6.1 Graph Density

Define edge density (with equation

5.6.2 Results

Show links between follower audience -> local network -> density.

5.6.3 Uses

From this data, it is demonstrable that various parameters can be generally successfully inferred from very basic user information.

5.7 Predictions From the User Graph

The final part of this paper focuses on the ongoing development of a method to predict the interestingness of a tweet based on the work in the previous sections. The prediction method, at a high level, compares the predicted retweet outcome of a given tweet to the number of times that tweet has actually been retweeted. If, for example, a tweet is simulated with the help of the model and produces a prediction of two retweets, but the tweet has actually been etweeted four times, then we can infer that this tweet is more interesting (at least, to a subset of users).

The tweet and user features we looked at earlier in this paper are very static, binary features, which do not take into account the actual content of the text of the tweet. Therefore, if a tweet is retweeted more than was predicted, then there is something in the tweet, such as a link to a particularly interesting article or a breaking news story, that makes it more interesting than the average tweet, with the same binary features, that was used to train the model.

In order to improve the fairness of the experiment, we wanted to ensure that the environment of the tweets (i.e. the user network they are propagated through) is the same as its real-life Twitter counterpart. We could then choose a user, which would become the source user, U_s, in the set U, and simulate that user's own tweets within their particular local network as described by the model above. This would then produce a retweet volume for this user's tweets, to which we could compare the number of times that tweet has *really* been retweeted.

5.7.1 Data Collection

Due to the exponential scaling properties of Twitter's social graph, it was infeasible to collect any more than two hops away from each user as a representation of that user's local network under the rate limitations of Twitter's REST API.

In particular, a single Twitter account running an instance of an application was allowed, at the time of these experiments, a maximum of 350 REST API calls per hour. One call would be required, for example, to obtain up to 5000 of the followers of a particular user (i.e. one follower hop from the user). An additional call would then be required to collect each of that user's follower's followers (in order to obtain the *second* hop from the source user).

Thus, a user who has 700 followers would require 700 API calls to collect that follower network, in addition to the one required to collect that source user in the first place, and would therefore take over two hours of collection. To collect the *third* hop from the source user would drastically multiply the number of required requests (even if there

is significant overlap between the followers) and the time needed.
If each of the 700 followers of the source user had, on average, 200 followers, then this would require the gathering of 700 x $200 = 140,000$ users, equating to more than 402 hours of data collection. Bearing in mind that this would only collect the network features for *one* user, it is clear to see how this is an impractical approach.

Luckily, in [15], we found that the vast majority of retweets occur *within* two hops of the source user (i.e. a path length of less than three), so we considered that the distance from the source user in each case would be sufficient.
In June 2012, the Twitter REST API was used to conduct a random walk through the social graph. For each user, we collected the most recent 300 tweets (including each tweet's metadata - particularly their retweet count) and their local follower network within two hops. We didn't collect the friend network, as we were only interested in tweets propagating outwards from the source user.
After processing that user, the walker chose a user at random from the present user's set of followers and made this the new current user from which to collect data for. If the present user, at any stage, does not have any followers, a list of previously accepted users is maintained and a follower is chosen from one of those instead.
The walker continued until the rate limit was met, at which time the current state was written to disk, and the walker waited until the rate limit was reset before continuing.
Generally, this resulted in, for each user, a set of up to 300 tweets (totalling to around 10,000 tweets in total) and the network in which these tweets were propagated within.
There was no need to collect any further data to train the regression, since we were able to re-use the trained model we used earlier.

5.7.2 Validating Results

Needed to validate results using human input. Machines themselves are generally unable to express human interests, so results need to be properly evaluated.

Crowdsourcing

Discuss crowdsourcing, its uses, how it is useful in this area. Talk about its history (with any references), and then about mechanical turk.
Mention mechanics of mechanical turk, how it is US only (but we used crowdflower - which autmatically handles submission to MT and several other crowd-sourcing services.

Preparing Validation Method

Set up questions (i.e. 5 tweets - choose most interesting and least interesting), give example of this.

In order to validate our prediction results, we ran a pilot user study in order to obtain some human input on the interestingness of each tweet. We compiled the tweet data into a set of questions which were submitted to Amazon's Mechanical Turk. Each question consisted of five tweets from our dataset and each Mechanical Turk Worker (MTW) undertook five questions. Each question asked the MTWs to select which tweet was the most interesting of the five, and which was the least interesting.
For consistency we ensured that at least three MTWs had answered each question. When selecting tweets to include in the Mechanical Turk questions, we excluded those which are '@-replies' - i.e. tweets which begin with another user's screen-name and typically form part of a conversation between two or more users. This meant that there were around 4,500 tweets in total in the questions.

Through using the model and simulating each user's tweets through their individual local networks we achieved around 86% accuracy in correctly predicting the number of times each tweet was retweeted.
The precision in predicting the *interestingness* of each tweet was around 30%. While this value is low, it does mean that in 30% of cases, a tweet that we predicted to be interesting was verified to be interesting by at least three MTWs all selecting one tweet

from a set of five. In addition, when simulating the questions by randomly choosing the most 'interesting' tweet of the five in each case, the performance was unable to near our precision even after several thousand iterations.

5.7.3 Improving This

Need offline methodology.
One route for this would be to try and infer a user's local network from a set of their immediate parameters, drawing on our earlier work suggesting that the Twitter network has the properties of a scale-free small-world graph. Through studying graph patterns, it is possible to make sensible inferences on the edges and nodes of a user's local network based on their follower count. From this, a graph edge density can be calculated, $d = \frac{|E|}{|N|(|N|-1)}$, for use in generating a scale-free network.
Since, for these preliminary experiments, we were only able to collect data from users with a more modest local network, the real and predicted retweet values were both relatively low, allowing more room for error. When simulating much larger local networks involving many more real retweets for each tweet, predicting interestingness, with some threshold value, may become more accurate and thus help improve the precision. The reason for this is that the retweet count of tweets that naturally get retweeted many tens, hundreds, or more times is likely to vary more with interestingness than those that are naturally only retweeted very few times.

In this chapter we aimed to carry out a study on the behaviour of propagation through different types of social graph structures and to introduce our ongoing work into predicting the interestingness of tweets from their retweet patterns.
Using a set of tweet and user features, we trained a regression model which we used to simulate a number of tweets through different network types. We produced a distribution of retweet volumes for each network type and confirmed that, with the same tweet features, different network configurations do indeed facilitate different retweet behaviours in terms of propagation spread. We were also able to compare our results

to data from Twitter to verify that Twitter's own social graph most closely resembles a scale-free small world graph.

We then finished by discussing how we used the trained model to simulate real networks from Twitter, along with the tweets that were passed through these networks, in order to try to predict how interesting a tweet is based on its retweet patterns. While we were able to often correctly predict the retweet outcome of a tweet, we found that more work would be required to improve the performance of predicting whether or not these tweets are truly interesting to users.

Chapter 6

Inferring Interestingness of Tweets based on Information Flow Through the Network

Introduction to online interestness inference.

Mention:

- How this chapter builds upon network stuff in previous chapter
- We hope to compare and contrast two better ways of predicting retweet volume *and* interestness
- What needs to be improved (speed, usability - more users with more followers etc.)
- Why do improvements need to be made?
- How is this useful, and how does first chapter relate to work done here?

6.1 Inference Generation Using Direct Predictions

Discuss:

- This method does not use a network or model individual user decisions
- Trained on a set of that particular user's tweets with the retweet outcome of integer type
- A new tweet modelled with the regression outputs a retweet volume prediction without having to simulate the Tweet's travels through the network.
- Discuss about the machine learning approach used (logistic regression and how it works)
- Talk about the 'binning' of retweet outcome volumes and its approaches (distribution dependent / independent, tables of precisions, etc.)
- Link 'retweet volume' to 'retweet group size'

6.1.1 Background

Stuff introducing this section and its differences to the things in the previous chapter. Whilst machine learning was used in the previous chapter... more in depth in this area.

6.1.2 Machine Learning

Explain about Machine Learning, its uses, techniques and how this is useful.
Talk about how it was used in previous chapter, but that more in depth here.

6.1.3 Features

List the features collected, how they were collected, when, etc.
Perhaps use a table listing user and tweet features.
Decision to bin the retweet outcome (more classifiers available, more accuracy, etc.)

6.1.4 Binning Strategies

Describe the different methods (linear, distributed 1, distributed 2), and their advantages/disadvantages, with examples showing the graph of what the bins look like.
Focus on the distributed 2 example, and why this is better.
Number of bins: explain how accuracy worsens as bin number increases.
'Requested' bin number not usually the same number as what is actually returned (due to large numbers of smaller retweet groups).
Pseudocode

6.1.5 Classification

Compared several types of classifiers (show table comparing accuracy, etc. of different types)
Explain that Bayesian Network is best (quick, accurate)

6.1.6 Training Results

Playing with Weka to improve the

6.1.7 Data Collection

6.2 Comparisons

- Comparison of two approaches
- Which is more accurate?
- Which is quicker?
- Which is easier?
- etc.

6.3 Summary

- Summarise comparisons - which method is overall better?
- How does it compare to offline method from previous chapter?
- Generally explain how the ‘winning’ method is good and accurate.

Chapter 7

Critical Assessment and Conclusions

7.1 Critical Analysis of Results

7.1.1 Analysis of Initial Research

- What was the use of initial research?
- Are the results sensible?
- How have the results shaped the further research?

7.1.2 Analysis of Final Results

- Have methods been able to sensibly predict retweet volumes?
- Have methods sensibly inferred Tweet interestingness?
- What might have worked better?
- Which parts were useless?
- Which parts helped develop other areas of research which may have provided further avenues of research ideas?

7.2 Further and Future Work

How can this research be taken further in the future?

- Use previous results to predict how far a tweet is likely to be retweeted (for advertising purposes)
- Useful for detecting the kind of messages that are likely to travel further
- As well as providing an interest level, the systems also predict sensible estimations on retweet volumes.
- Perhaps useful for measuring the spread of rumours.

7.3 Conclusions

7.3.1 Summary

Summarise events and processes covered, reiterate what the point of the work was and how each part of the work covered relates to that.

7.3.2 Contributions

Restate the original contributions (from Introduction section). Explain the ways in which the work done relates to the projected contributions, that it is novel and useful.

Appendices

Source code, further diagrams, ideas, etc.

Bibliography

[1] 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.

[2] Paavo Arvola, Jaana Kekäläinen, and Marko Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13:460–484, 2010. 10.1007/s10791-010-9133-9.

[3] C Castillo, M Mendoza, and B 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.

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media*, May 2010.

[5] M J Chorley, G B Colombo, S M Allen, and R M Whitaker. Better the tweeter you know: social signals on twitter. In *International Conference on Social Computing, ASE/IEEE*, pages 277–282, 2012.

[6] A. Clauset, C. Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *ArXiv e-prints*, June 2007.

[7] 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, pages 3–3, Berkeley, CA, USA, 2010. USENIX Association.

[8] O Hein, M Schwind, and W König. Scale-free networks. *WIRTSCHAFTSINFORMATIK*, 48:267–275, 2006.

[9] L Hong, O Dan, and B 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.

[10] A Java, X Song, T Finin, and B Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[11] 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.

[12] H-K Peng, Z Jiang, P Dongzhen, Y Rong, and Z Ying. Retweet modeling using conditional random fields. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 336 –343, 2011.

[13] Bongwon Suh, Lichan Hong, P. Pirolli, and E.H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 177 –184, aug. 2010.

[14] I 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, New York, NY, USA, 2011. ACM.

[15] W Webberley, S Allen, and R Whitaker. Retweeting: A study of message-forwarding in twitter. In *Workshop on Mobile and Online Social Networks (MOSN), IEEE*, pages 13 –18, 2011.

[16] 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.

[17] Dejin Zhao and Mary Beth 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.

[18] J Zhu, F Xiong, D Piao, Y Liu, and Y Zhang. Statistically modeling the effectiveness of disaster information in social media. In *Global Humanitarian Technology Conference (GHTC), 2011 IEEE*, pages 431 –436, 2011.