On-the-fly spam-detection on Twitter

Mourjo Sen
sen.mourjo [at] gmail.com

Under the guidance of
Arnaud Legout, Maksym Gabielkov
{**arnaud.legout**, maksym.gabielkov} [at] inria.fr

Master Thesis

Ubinet Master of Computer Science
Department of Computer Science
University of Nice Sophia Antipolis
France

August 2015

On the fly spam-detection on Twitter

Mourjo Sen

Under the guidance of
Arnaud Legout, Maksym Gabielkov

Submitted for the degree of Master of Science
August 2015

Abstract

Twitter has become an integral part of our lives – so much so that public opinion is often estimated by trending topics on Twitter. From results of democratic elections to protesting against government policies – anything can trend on Twitter. While this has many positive effects on society (like gathering funds for natural calamities or promoting educational activities), there is also the dark side of social media. Since trends attract a large audience in a snowballing-like effect, there is often a huge incentive to falsify or manipulate trends on Twitter (or other popular social networks). From companies to politicians – many people can benefit from trending topics about themselves or their products. It is in this context that being able to differentiate between manipulated tweets (or spam) from authentic tweets becomes extremely important. This work aims to provide a filtering mechanism, by providing an automatic spam detection technique for tweets on Twitter, thereby reducing the chances of fabrication of public opinion on Twitter.

Acknowledgements

I would like to thank my supervisor Dr. Arnaud Legout for providing useful research direction at every step and his help and guidance during the course of this internship. I am grateful to Inria Sophia Antipolis in general and Dr. Legout in particular for providing high-performance hardware without which analyzing the huge datasets of Twitter would have been extremely difficult and cumbersome. Furthermore, I would like to thank Maksym Gabielkov for his help especially with the technical challenges of the internship. I would also like to thank the other members (Anuvabh Dutt and Anastasia Kuznetsova) in the Diana team at Inria who are also working on the analysis of Twitter for their insight and interpretations of the results. Lastly, I would like to thank my parents for their constant love and support for every venture in my life.

Contents

Chapter 1

Introduction

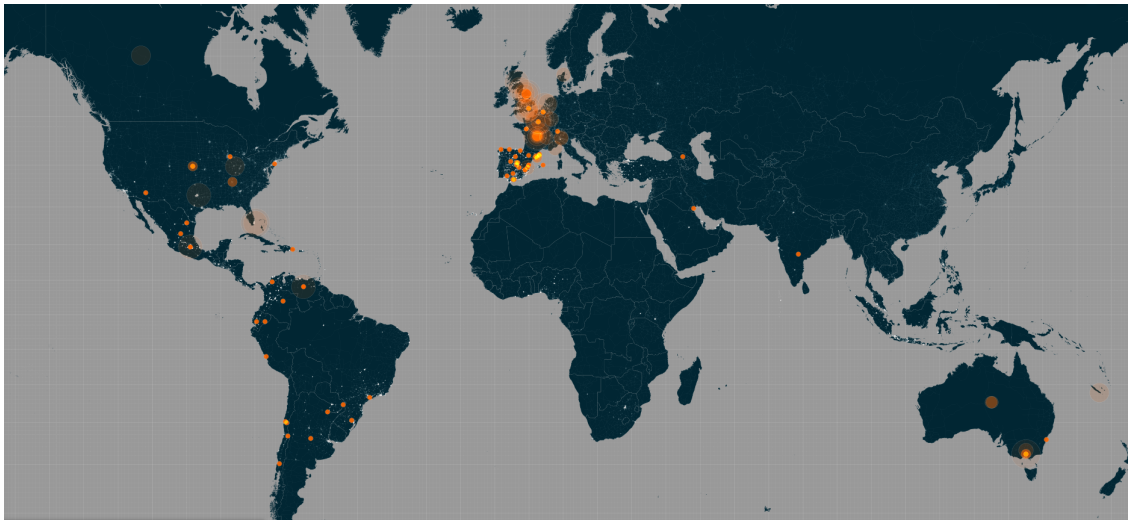
Twitter is a very popular online social networking and microblogging service that enables users to send and read short 140-character text messages, called "tweets" [1]. Most of these messages are public and can be viewed without even creating a Twitter account. Twitter has more than 537 million users and about 255 million monthly active users. Twitter provides a huge platform for many aspects of social life today – from assessing public opinion to popularizing protests. While Twitter is very significant today as a means for quick information propagation, a serious question arises if such strong influence on society can be manipulated for wrongdoing. This is what our work aims to explore.

1.1 Motivation

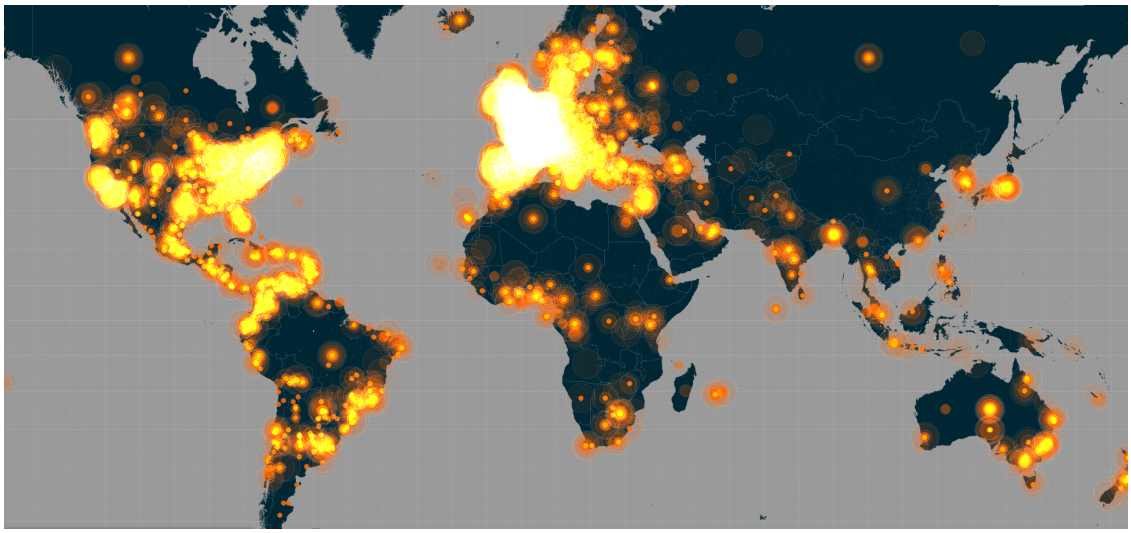
There are about 500 million tweets published every day [2]. It is easily possible for someone to pretend to be someone else on the Internet (impersonation) and/or publish fraudulent information with malicious intent or for unethical professional gain, or simply to create rumors [5]. Therefore, it is very important for any information (here tweets) to be proved authentic and trustworthy. There have been instances of social media hoax that have had serious sociological connotations in recent years. From creating rumors of actor Morgan Freeman's death [22] to a teenage girl's kidnapping hoax [23] – all have been fueled by the *quick, unchecked* and *effortless* diffusion of information in social networks in the wild. In fact, Twitter has also been used for more serious crime like stock market fraud [25] and other illegal financial activities. With more than five hundred million users on Twitter, it is almost impossible to manually verify the identity of every user who signs up on Twitter and it is even more difficult to keep track of users who tend to spread information of questionable authenticity, unknowingly or deliberately. Therefore we need to identify trustworthy users on social networks like Twitter, and more importantly be able to separate spam from legitimate tweets.

Figure 1.1: *Twitter France published their data on the "Je suis Charlie" campaign [26] – as stated by them, "2,100,000 tweets showing solidarity on* `#JeSuisCharlie` *published today". Not only does this figure show that Twitter represents public opinion, but it also shows that Twitter can be used to popularize a movement or topic, that is, Twitter makes it possible to make a local movement a worldwide phenomenon in 24 hours (see figure 1.2). It is because of this high influence of Twitter on the real world that being able to separate legitimate and authentic tweets from spam and fraudulent tweets becomes very crucial.*

Being an information *broadcasting* mechanism, information on Twitter can quickly become viral. This has been manipulated in the past for personal and professional gains. The strong incentive to manipulate perception of public opinion lies in the fact that Twitter has an ever-increasing impact on society. For example, the "Je suis Charlie" campaign was started on Twitter [30] and was fueled by its popularity on Twitter. Within a matter of hours, the `#JeSuisCharlie` became a worldwide trending topic on Twtiier (see figures 1.1 and 1.2). Social media analytics (from Twitter feed) are used more and more to assess general public opinion, which has made it financially and/or personally very viable to be able to "hack" trends on Twitter. For example, one may be able to wrongly convince people that one's product is highly talked about on Twitter. This would cause increase in the product sales because it would seem to be a really popular product, while in reality, it was just a manipulation of the trends on Twitter. Because of such strong incentives to manipulate trends on Twitter, it is of utmost importance to be able to filter out such opinion-altering fraudulent spam tweets from the authentic real tweets and be able to do so in a robust way in real time.



(1.2.1) *Geo-localized tweets containing* `#JeSuisCharlie` *at 1:47 pm (CET) on January 7, 2015.*



(1.2.2) *Geo-localized tweets containing* `#JeSuisCharlie` *at 10:35 pm (CET) on January 7, 2015.*

Figure 1.2: *The popular slogan "Je suis Charlie" for the Charlie Hebdo incident first appeared on Twitter [30]. The campaign's slogan* `#JeSuisCharlie` *spread across the world in just a few hours with the help of Twitter [28]. Twitter can be used in this way to popularize campaigns worldwide. But there may be serious consequences if such viral campaigns were fraudulent with or without malicious intent.*

1.2 Background

Before we talk about solving the problem of differentiating authentic tweets from spam tweets, we have to state the context of the problem and some background information that will be required when we explain our method. We explain such notions in this section.

1.2.1 The Problem in Twitter Today

Twitter defines spam as tweets that violate the rules of Twitter as stated on their support page (`https://support.twitter.com/articles/18311#`). As Twitter states, the major problem in detecting spam is that it cannot be defined accurately [38], and even Twitter is struggling to first define and then block spam. We talked with researchers from Twitter, and also with a social media analytics company dealing directly with Twitter, and from both of these interactions we have found that Twitter is not being able to get rid of spam from the social network. Finding and blocking spam on Twitter is a problem even for Twitter and they are also actively trying to prevent it [40]. Most existing solutions use semantic analyses and machine learning techniques, which work for some cases but fail for others, as we will show in the following chapters. Our goal is to find a robust, on-the-fly spam detection method that is deterministic and performs well in all situations. Out method focuses less on the content of the tweet and more on the authenticity of the user who publishes or retweets the tweet.

1.2.2 Twitter Terminology

Twitter is a microblogging service. This means that on Twitter, users write short 140-character long messages called "tweets", instead of long posts as is the case for traditional blogs [32]. These 140-character messages, or tweets, are the building blocks in Twitter. Tweets are responsible all activity on Twitter – from running very successful campaigns and obscure, unread content to spamming. Apart from plain text, tweets can contain special tags which link to other users or topics. Tweets can be redistributed by users who like them and by doing so, some tweets become highly popular and influential. These are explained in the following sub-sections.

Types of Relationships in Twitter

There are two types of relationships between users in Twitter – follower and following. When a user Charlie follows another user David, Charlie receives all tweets composed (and retweeted) by David, but David, because he is not following Charlie, does not receive any of Charlie's tweets. In this scenario, Charlie is a *follower* of David and Charlie is one of David's *followings*. The follow relationship is not

symmetric and therefore a person following another person does not mean that the converse is true. In other words, the links on the Twitter social graph are directed. This directed relationship has caused Twitter to have many unique properties which are discussed later (in section 1.2.4).

According to Twitter's support page [27], a user Bob following another user Alice means **(a)** Bob subscribes to Alice's tweets (see figure 1.3), **(b)** Alice's updates will appear in Bob's home tab, **(c)** Alice will be able to send direct messages to Bob. Twitter defines followers as "people who receive your Tweets". If Bob is a follower of Alice, **(a)** Bob will be listed in Alice's followers list, **(b)** Bob will see Alice's tweets in their home timeline whenever they log in to Twitter, **(c)** Alice can start a private conversation with Bob (via direct messages).

Hashtags and Mentions

In Twitter, users are assigned a unique username. A user can refer to (or tag) another user in her tweet by using her username and appending with it an "`@`" symbol, like "`@username`". Such references to users in tweets are termed as mentions. There is another set of tags that can be used in a tweet, called hashtags. Hashtags are denoted by a "`#`" symbol followed by a set of alphanumeric characters (without a space), for example, "`#hashtag`". Hashtags are created dynamically and usually refer to a topic. Hashtags help locating tweets with a similar topic or theme [33]. For example, `#JeSuisCharlie` was one of the most used hashtags in the history of Twitter [30], signifying the world's solidarity with the Charlie Hebdo incident in France. The presence of the hashtag `#JeSuisCharlie` in a tweet suggests that the tweet was written in context of the Charlie Hebdo incident. In most cases, a highly popular topic has one designated hashtag.

Mentions and hashtags are integral parts of a tweet and the presence of such tags can change the fate of the tweet, in terms of visibility. When hashtags or mentions are present, users who are not following the author of the tweet can also find this tweet by searching for the hashtag or the user mentioned in the tweet. Moreover, by clicking on trending hashtags shows tweets that contain the hashtag. So, if a popular hashtag is mentioned in a tweet, it is likely to be seen by more users than just the followers of the author. In fact, one of Twitter's own blog articles states that tweets with hashtags and mentions receive more visibility than those without [34].

Retweeting, Favoriting and Replying

Twitter provides three major ways for users to publicly interact with tweets, namely retweeting, favoriting and replying to a tweet. The three modes of public interaction are quite unique and have their own benefits and drawbacks. In this section we will define each of them and later (see section 2.2.2) go into the discussion of which one is more suited to our spam detection method and why.

Twitter provides a feature called "retweeting", using which a user may redistribute

a tweet to her followers. Twitter defines retweets as reposting of an already-existing tweet, which was authored by someone else [39]. For every retweet, there is a user other than the author who does the retweet – we call that user the "retweeter". Retweeting is similar to sharing a post on Facebook – it tells one's followers that one recommends a tweet. For example, if a user Bob receives a tweet from Alice (because Bob is following Alice), and Bob decides to retweet this tweet, then all of Bob's followers will receive this tweet in the same way as they would have if it was a tweet authored by Bob, the only difference being that the author of the tweet will still be Alice and not Bob. They will receive it even if they themselves are not following the original author Alice. The act of retweeting and the properties of the retweeter have very significant connotations, which are described later in this report (see section 2.2.2).

Another way for a user to interact with a tweet is to favorite it. Favoriting a tweet can be seen as a way of bookmarking the tweet to come back to it later, and to get alerts regarding the tweet. Unlike the other two modes of interaction, favoriting does not engage other users directly – that is, favoriting a tweet does not create any new content nor does it redistribute existing content. It is simply a way to show one's appreciation for a tweet and that one may come back to this tweet later in the future.

The third way for a user to interact with a tweet is to reply to it. When a user replies to a tweet, it is similar to commenting on a Facebook post or blog article. The act of replying is a way for a user to share her views on what a tweet says. She may be sharing her personal opinion on the content of the tweet, asking a question regarding the tweet, or answering a question that was asked in other comments or the tweet itself. She could also be replying about a separate topic altogether. Replying to a tweet starts a conversation on its own, which may have little or no connection to the original tweet's content.

Trends on Twitter

Topics that are very popular in real time are called "trends" on Twitter. Trends were initially created from the most used hashtags, but now Twitter also shows keywords as part of the trends. Twitter trends are dynamic and change depending on the tweets being created at real time. Trends can also vary depending on the location of the user and who this user follows [36]. The way Twitter finds trends from published tweets has not been disclosed by them but they have stated that trends are generated by their proprietary algorithm, which tries to find the topics that are being mentioned currently in tweets more than they were being mentioned in the past – that is, their rate of occurrence in the tweet stream is increasing. They have explained that a topic can start trending when the number of times it is being mentioned increases suddenly and dramatically. This also means that highly popular all-time topics may not be included in the trends (since their rate of citations is not increasing). In other words, Twitter trends prioritize novelty over popularity to determine trends.

According to Twitter, trends were designated to help the user find breaking news and hot emerging topics on social media in real time. Trends are showcased in the Twitter homepage after signing in, so the trending topics/hashtags are given the maximum visibility. Thus trends may change the perception of the individual user, in the sense that the user may be more inclined to tweet about (or find out more about) topics that are extremely popular at that moment [35]. As the Twitter user community as a whole has the ability to change trending topics, so do trends with regard to convincing individual users to tweet about a trending topic. So trends are more likely to get even more tweets, in a snowballing-like effect.

Twitter trends have affected the world so much that it caused the English dictionary to include a new meaning for the word "trend". One of the meanings of the word "trend", as currently stated in the Oxford English Dictionary, is *"a topic that is the subject of many posts on a social media website within a short period of time"*.

1.2.3 Verified Users

Some users on Twitter have been manually verified to be authentic and genuine by Twitter following administrative processes of cross-checking their profile with their identity. Mostly verification is done for public personalities and celebrities who are famous. According to Twitter, "Verification is currently used to establish authenticity of identities of key individuals and brands on Twitter" [6]. Twitter certifies that these verified users are trustworthy. Only a very small fraction (0.01 %) of users is verified by Twitter, and it leaves the remaining 99.99 % of users, whose authenticity and legitimacy cannot be guaranteed by the lay man. We will use the certified authenticity of verified users and extrapolate it to find more users who can be trusted as well.

1.2.4 A Model for Real-Life Information Propagation

As Twitter evolved over time, celebrities and famous personalities started using Twitter [29] and because Twitter was then the first platform to directly interact with authentic celebrities and famous personalities, they started being followed by a lot of users. Thus, there emerged two types of users – those who are followed by many people and those who followed many people. Since the celebrities' tweets reached a lot of people on account of the large number of followers, the celebrities started becoming "information producers" and others became "information consumers". The root cause for these different classes of users is that the relationships on Twitter are asymmetric. This has caused Twitter to be considered as a news medium as well as a social networking medium [3]. This type of information production/consumption makes Twitter different from other social networking sites like Facebook, which primarily uses undirected links between users resulting in the symmetric information flow – where "friends" are both producers and consumers of information at the same time. Of course, symmetric information flow can be achieved on Twitter when two

Figure 1.3: *The follow relationship is inverse to that of tweet subscription. When a user follows another user, the first user receives all the tweets from the second.*

users follow each other. Therefore, Twitter is more general and is closer to how information propagates in real life [4].

It has been shown by H. Kwak et al. that the act of following is mostly not reciprocated [3]. Rather, there are a few users who have many followers and can therefore directly reach a large audience, which is what the celebrities trend started during the early years of Twitter. This aspect can be thought of as the "fan relationship" in real world. That is, celebrities have a lot of fan following, but ordinary people have few. The relationships found in Twitter model this asymmetric relationship, as well as the symmetric friendship relationship (when users follow back each other). It is because of the presence of these two types of general relationships in Twitter that we believe that it can be a good model of social networking in general and hence we chose to study the aspect of trust and spam in Twitter.

1.3 Related Work

Our work is closely based on a set of users that we call trusted. Thus, the core idea revolves around correctly identifying trusted users. In this regard, a significant amount of work has been done on how to find trusted users in a network and ensure network security. We focus on finding trusted users on social networks like Twitter. Techniques such as chi-square distributions [19], SVM-based methods [13, 14] and machine learning techniques have been used to estimate trust [15]. TwitterRank [16] is another approach which is based on the PageRank algorithm [31] to find influence of users on certain topics. Pal et al. [20] have used features of the Twitter graph and analysis of tweets to deduce if a user has significant contribution to a topic, which gives them a level of trust towards that user in that field.

Twitter has its own "who-to-follow" service which uses information from a user's profile to find popular users with the same field of interest. According to Twitter, these suggestions are based information like email and/or phone contacts, contacts of contacts and patterns detected from history of behavior on Twitter [18]. Of course, the users suggested by Twitter have either personal relation (like phone contacts)

with the user or have had a significant impact on Twitter social network and can thus be assumed to have a trusted status.

Naveen Sharma et al. [9] have created a "who-is-who" inference service (deployed on the Web at `http://twitter-app.mpi-sws.org/who-is-who/`) using Twitter data to deduce particular attributes about users. The information deduced from Twitter by the "who-is-who" service not only uses information in users' Twitter profiles but also their areas of interest and expertise. Moreover, it computes popular view about users on Twitter.

Most approaches before Naveen Sharma's work concentrated on information about users from content that is produced by the user themselves, like profile information, biography, and tweets [17]. But such information may not be accurate in describing whether a user can be trusted or not because such content can easily be manipulated by users with malicious intent. Naveen Sharma et al. therefore used a different approach to analyze the online personalities of users on Twitter. They used Twitter lists, which are groups of Twitter users created by a particular user to manage tweets related to similar topics. According to Twitter, one can create one's own lists or subscribe to lists created by others. A list timeline shows a stream of tweets only from the users on that list [11]. Many users use lists to categorize the people who they want to follow. Thus, if a user appears on many lists, it would suggest that the user is popular.

The work of Naveen Sharma et al. was to examine the meta-data of Twitter lists in which a user appears. That is, to study how popular a user is, and more importantly to study in what field a user is popular in, Naveen Sharma et al. used the information about the lists themselves to gather a user's field of popularity, and not just popularity as a whole. This gives more insight because social networks are highly clustered and people are often only popular among people having similar interests. Attributes of users found using this method include their field of popularity or expertise, interests, professional details, "known for" attributes. For example, Lance Armstrong may state in his profile that he is a noteworthy sportsperson but details about events like his participation in Tour de France and association with cancer can be deduced by studying the lists he appears in. But whether this correlates to the issue of trust has not been studied.

In their study, Naveen Sharma et al. divided popular users, found using their technique, into three classes based on the information gathered using their technique: well-known users, news and media users and US senators. They found a set of users whom we call experts, who have a high influence on Twitter because they are related to a certain domain where they are popular. Amongst these users, there are users who are experts on very niche topics, such as robotic space exploration, and stem cells. They have had a considerable impact on the Twitter social graph. We will use the set of experts in assessing our trust metrics. Analysis of the results of this study showed that very popular users, i.e. users with a high number of followers, are mentioned in lists quite frequently. The number of mentions in lists falls quickly as the number of followers decreases.

There are users in Twitter who have been verified by Twitter and constitute a small fraction of the total population of users in Twitter. Verification is only initiated by Twitter and it cannot be requested by a user. It is typically performed for celebrities and public figures, including notable businesses. We consider a verified user to be trusted (relative to an unverified one), since to earn verified status, they must use their real-life identity and be sufficiently well-known by the general public for Twitter to consider that there is a risk of someone impersonating them. As stated by Forbes.com, several businesses have been impersonated, with impostors posting tweets that threaten to malign their business's reputation [1]. The verified status is thus a mechanism to uphold the integrity of the user's identity. Therefore we consider a verified user to be trusted.

After the work done by Naveen Sharma et al., the structure of Twitter has gone through a few changes. Naveen Sharma's work was based on Twitter data from 2009, collected by Kwak et al. [3]. As studied by Maksym Gabielkov et al. [4], Twitter has changed since then.

1. The 2009 dataset (along with other previous data sets) was not exhaustive/ complete and there were subtle properties that were not visible in the 2009 dataset [4].

2. There was a major change in the Twitter graph in 2009-2010, when a large number of celebrities joined Twitter and its popularity started increasing manifold. Celebrities are not interested in following too many people, but they have a high number of followers. This changed the properties of the Twitter graph in 2009-2010. These changes were not reflected in the 2009 dataset [3] used by Naveen Sharma et al.

3. There has been a tenfold increase in the size of Twitter since 2009. Twitter today is one of the most popular social networking sites on the Internet [7].

Thus, for this work we have to use more recent datasets. The dataset collected by Maksym Gabielkov et al. [8] represents Twitter of 2012. They used distributed crawling with 550 PlanetLab nodes to collect the dataset. This dataset contains the entire Twitter graph of 2012, i.e., every user on Twitter and their connections. This dataset is significantly more comprehensive, exhaustive and more representative than the 2009 dataset. We used this dataset in the work.

Chapter 2

Methodology

Having defined the required Twitter terminology, in this chapter, we talk in details about our spam detection method. First we talk about the dataset and tools used to deduce the method we propose, and then we will go into the details of the algorithm to detect spam tweets.

2.1 Datasets and APIs used

Our method required two types of data from Twitter – firstly, about the users, and secondly, about the tweets. For the user data, we first tried different methods that we thought would be applicable in quantifying the trustworthiness of users and then to identify spam. We used a partial dataset of the Twitter social graph in 2009 [9] collected by Naveen Sharma et. al. We then used a complete Twitter user dataset of 2012 collected by Maksym Gabielkov et al. [8]. This was used mostly to find trustworthy users in Twitter in 2012. We then partially collected data from Twitter in 2015 to see how the our method handled the changes in Twitter time from 2012 to 2015. We used the 2015 dataset to show that our method can adapt to changing dynamics of the Twitter social network over time. To collect data from Twitter in 2015, we used Twitter's public REST and Streaming APIs, and had to work with the rate limits imposed by Twitter [24]. We used a random sampling technique to collect some portions of the graph in 2015. The reason we could not collect the full dataset of Twitter in 2015 is because of the huge size of Twitter, and the rate limitations imposed by Twitter's REST API which made it practically impossible to crawl the entire Twitter graph without using any specialized infrastructure like NEPI [21]. For obtaining the tweet data, we first accessed the public API of Twitter and collected data over various time intervals. The public API of Twitter only gives access to a random one-percent sample of all tweets being created in real-time. Because the data from the public API was not complete, we used a complete dataset of tweets collected over three months obtained from a social media analytics company called Vigiglobe [45], which is one of the eighteen startups worldwide to have direct access

to Twitter's data. These tweets were collected mainly by searching for related keywords to Microsoft, like "Surface", "XBox", "Lumia", "Windows" etc. We had to use this third-party dataset because it was complete and all the tweets published during the time of collection are present in the dataset, which was not possible with the public API.

2.2 The Spam Detection Method

Our method to detect spam is divided into two main stages – first, we profile users using a metric called the "trust score" (higher a user's trust score, more trustworthy the user is) and then we use this trust score to quantitatively estimate the quality of users (the trust score) who are associated with a given tweet. A tweet is then given a score of its own, based on the type of users associated with the tweet. The spam detection then works by selecting a threshold on the tweet score, and if it is above the threshold, it is considered as not spam. The details on both stages are given in the following sections.

2.2.1 The Trust Score

The first goal of this work is to define a trust score for every user such that it indicates how trusted that user is – because once we have profiled a user's trustworthiness, the type of users associated with a tweet will give us the tweet's confidence from the community of Twitter users, which can then be extrapolated into a spam classification mechanism.

Though "trust", by essence, is a subjective notion and is open to interpretation, for this work, we define being trusted as the likelihood of being who one claims to be, that is, judging by a person's position on the Twitter graph, how probable it is that she is who she claims to be. By using this notion of trust, we hope to identify legitimate users on Twitter, who create trustworthy (i.e., not spam) tweets. To quantitatively estimate the inherently subjective notion of trust, we use a metric called the "trust score", which is based on the number of verified followers a user has (see figure 2.1). More number of verified followers means more trusted the user is. Any user with more than one verified follower is included in the "trusted set" of users. The threshold to demarcate membership in the trusted set of users can be increased for a more strict spam detection method, which we will come to later. We would like to make a clarification here that the set of verified users does not overlap the set of trusted users. Even if a verified user has a verified follower, because he is already verified, he is not included in the trusted set. Therefore, these two sets are distinct and disjoint. The set of verified users are found by Twitter by a manual administrative process whereas the trusted set is found by us in an automated process. The trusted set is much more inclusive than the verified set, which only contains 0.01% of all users on Twitter.

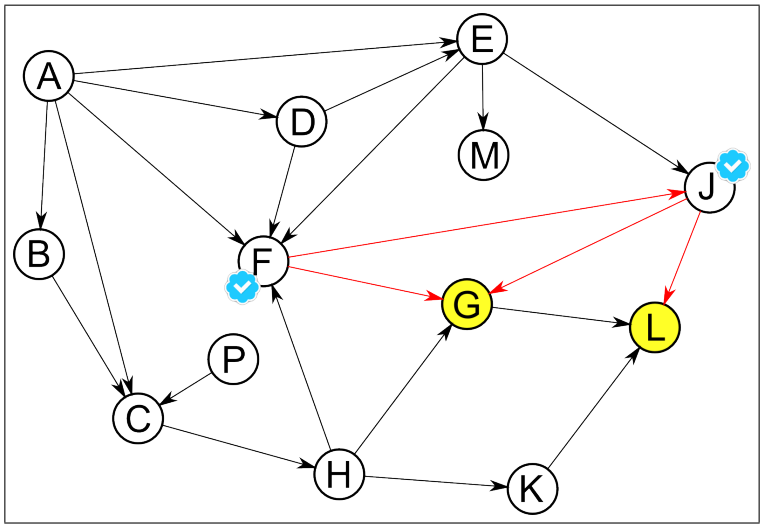


Figure 2.1: *A graph similar to the Twitter graph, with two verified users F and J. A link in this graph from X to Y means that X follows Y. The outgoing links (in red) from the verified users are used to find the trusted users. G and L are the trusted users with a trust score of 2 and 1 respectively. So G is more trusted than L. Verified users are not included in the trusted set.*

Since verified users have been certified by Twitter to be trusted and legitimate, we assume that a verified user will not follow someone who is known to be a fraud or an impersonator. Trust is imparted to a unknown user Bob by a verified user when the verified user decides to follow Bob. Imparting of trust is seen in other fields as well – consumers are more likely to believe what someone they trust says about a product rather than a marketing promotion of the same product. The notion of trust score is an extension of this – we know the verified users to be trusted, and when they follow someone, it imparts a value of trust onto the followed user.

The trust score is primarily important for users who have not been verified by Twitter as it gives an idea of how authentic a user is likely to be. Thus, with the help of the trust score, we get three categories of users in Twitter as explained in the next section.

Classification of Users Based on Trustworthiness

We divided all users into three categories based on the trust score of the users. The verified users are of course excluded from the trust score calculation because we are using the verified users as the starting point for the trust score calculation. Apart from the verified users, we have two new classes of users as explained below.

1. **Verified users**: As stated in the previous chapter, the users who have been manually verified by Twitter constitute 0.01 % of the total population of users in Twitter. Verification is only initiated by Twitter and it cannot be requested by a user. It is typically performed for celebrities and public figures, including notable businesses. We consider a verified user to be trusted (relative to an un-

verified one), since to earn verified status, they must use their real-life identity and be sufficiently well-known by the general public for Twitter to consider that there is a risk of someone impersonating them. As stated by Forbes.com, several businesses have been impersonated, with impostors posting tweets that threaten to malign their business's reputation [37]. The verified status is thus a mechanism to uphold the integrity of the user's identity. Therefore we consider a verified user to be trusted.

The verified users of 2015 were collected by traversing the Twitter graph and finding all the users who are followed by the Twitter account with the screen name `@verified` (`https://twitter.com/verified`). This account, while itself being verified, follows all other verified accounts on Twitter. As the description says on the `@verified` account, the `@verified` account follows "accounts verified by Twitter". We found that the percentage of verified users in 2015 has increased from 2012 by 1.6 times, and this suggests that Twitter is actively verifying users to increase the proportion of authentic users in Twitter. Therefore, having more percentage of users certified to be verified and thereby making the social network less susceptible to spam is of concern even to Twitter themselves.

2. **Trusted users**: These are the users who have a certain number (trust score) of verified followers. We consider these users as trusted because the verified users impart trust onto these users by the act of following them. We use this intuition to find the trusted users, and we will confirm this intuition in the next few sections. The trusted users (with trust score $>= 1$) constitute more than 6 % of users of Twitter, which is a larger set than the verified users, which makes it more representative than the set of verified users. The trusted set of users contain only unverified users, i.e., if a verified user is followed by other verified users, he/she is considered as a verified user, and not as part of the trusted set of users. The trust score defines how conservative the trusted set is. At the lowest value of 1, the trusted set contains all users that have verified followers, whereas at a trust score of 10, only users with at least 10 verified followers are considered trusted. Increasing the trust score dramatically reduces the size of the trusted set as depicted in figure 2.2.

3. **Other users**: These users constitute the rest of Twitter – that is those that are neither verified nor trusted. We cannot say much about these users with our method, but these users definitely include spammers (since they are not part of the trusted and verified set of users). We will show that the set of trusted and verified users is representative enough for our spam detection method.

The above classes of users have been created based on the expected amount of trust other users have in them (deduced from the Twitter social graph). But this does not indicate if the amount of trust affects their presence on Twitter – in other words, this classification does not tell us if the trust value of users has any impact on the popularity of the users. In figure 2.3, we see that verified users are highly popular and trusted users are more popular than other users. By being "popular", we mean having many followers. Thus, from figure 2.3, we can conclude that the amount

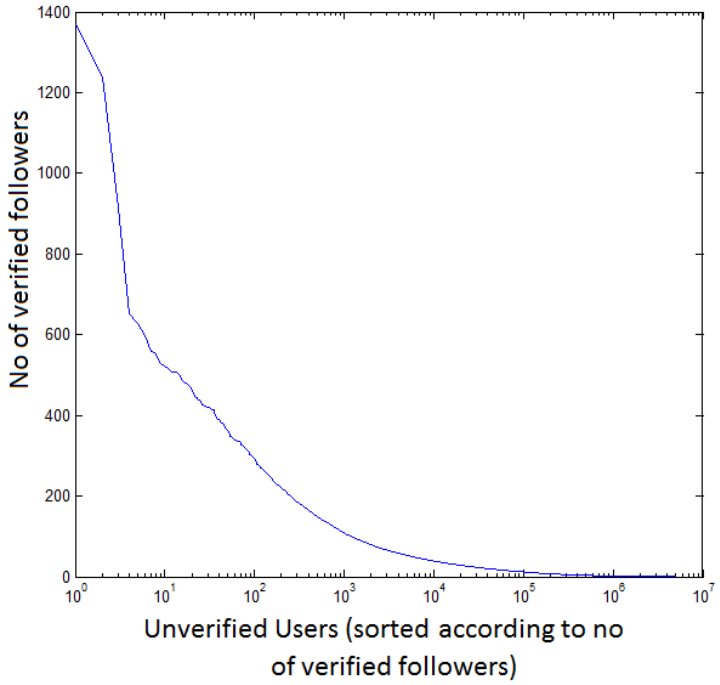


Figure 2.2: *This figure shows the number of trusted users in Y axis with increasing number of verified followers in X axis.*

of trust also agrees with the popularity of a user – verified users are more popular than trusted users, who are in turn more popular than other users. This serves as a validation in favor of our trust score metric and classifying users based on the trust score.

Justification of the Trust Score as a Metric for Trust

As we stated in the previously, only 0.01 % users are verified on Twitter and they are considered to be authentic. And since verified users are guaranteed to be of authentic identity, therefore, if an unverified user has many verified followers, it would suggest that this user is more trusted than other users who have no verified follower at all. This is the intuition behind the trust score, and in this section we will justify this intuition.

Verified users (0.01 %) are mostly celebrities and famous personalities. They tend to follow people selectively, although they often have a large number of people following them. We use this selective following characteristic of verified users as a checkpoint for estimating how trusted other users in the system are. Since verified users follow back only a small number of users, we can postulate that when any user has a high number of verified follower count, it means that the user is more trusted than one with fewer or no verified followers. This postulation is justified by the fact that if we only consider only verified users to be trusted, a person with a high number of verified followers shows that that person's activities (tweets) are followed by many trusted users, indicating that he himself must also be trustworthy.

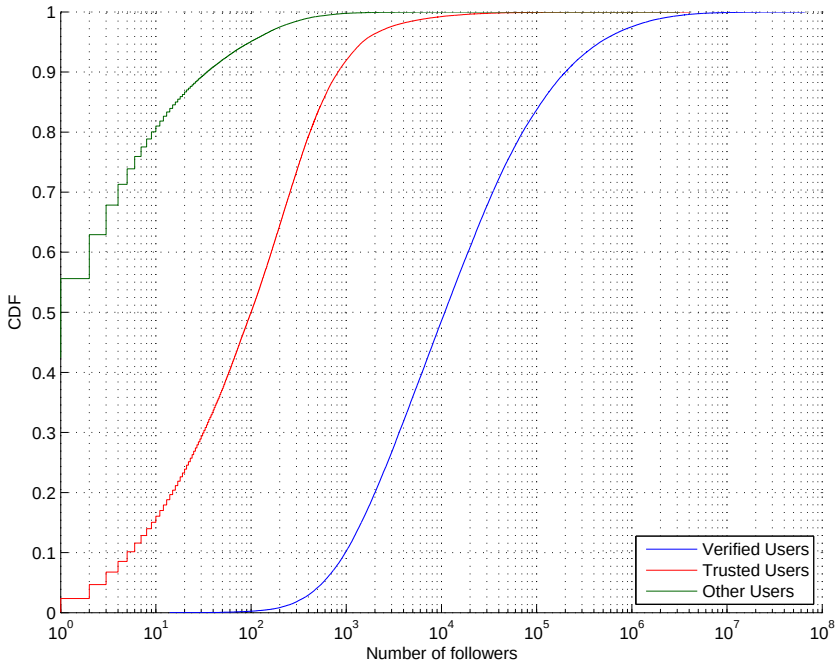


Figure 2.3: *This figure shows the CDF plot of the number of followers of verified users, trusted users and other users. This indicates that the trusted users are popular and are therefore followed more than other users. The popularity of trusted users may also suggest the amount of "faith" Twitter users have in trusted users – verified users have the highest popularity, then the trusted users and then the other users.*

When a user Alice follows another user Bob, it usually implies a level of trust that Alice has in Bob. Generally, people follow those who they personally know or are public figures, and no legitimate user is likely to follow a known spammer or a user known to be impersonating someone else. The quality of followers of a user thus provides a measure of trust. In fact, even Twitter provides a "followers you know" feature when a user visits another user's profile. The idea behind this feature is if many people who a user follows (and hence believes to be authentic) also follow this user, then there is a chance this user is also authentic – that is, Twitter also tries to express users' authenticity by the quality of followers (see figure 2.4). But the question is how to estimate the quality of the followers of a random user, or how to figure out which follower is legitimate. Since we know for certain that verified users are trusted, we can estimate the quality of followers by the number of verified followers a user has. Legitimate users impart trust onto those they follow. Since verified users are considered to be legitimate, users with a high number of verified followers are also expected to be trusted.

To have a concrete evidence in favor of the trust score, we compared the set of verified users of 2015 to the trusted set of 2012 and we found that 60 % of the new verified users of 2015 are from our trusted set of 2012 that we calculated (see figure 2.5). That is, using our trusted set (of 2012), we could anticipate which users are going to be verified by Twitter in the future (in 2015). The trusted set of 2012 consisted of just 4.6 % of the entire Twitter graph, and to find 60 % of the new verified users from that 4.6 % shows that our method is quite efficient in finding authentic users and

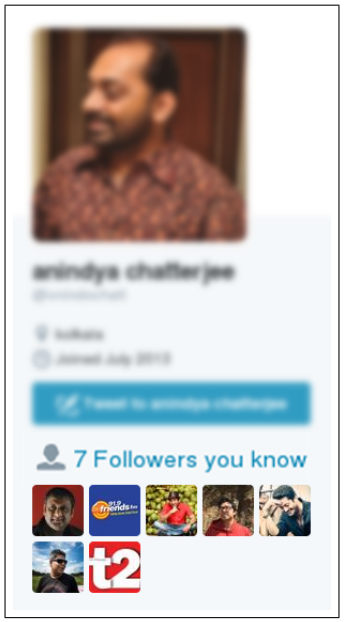


Figure 2.4: *The figure above shows a typical Twitter user's profile. Twitter's "followers you know" is an example of how important the quality of followers can be. When a user David visits the profile of another user Ethan, Twitter shows the people who both (i) follow the user Ethan and (ii) are followed by the user David. That is, if Alice → Bob → Charlie (i.e., Alice follows Bob and Bob follows Charlie), and Alice visits Charlie's profile, Twitter will show Bob under the "followers you know" tab. Thus, even Twitter encourages users to assess other users' Twitter profiles by the quality of followers they have. In this case, the quality of followers is measured by a transitive follow relationship.*

Twitter's verification process (and therefore their definition of authenticity) agrees with our trust score. Thus, our method not only strongly correlates to Twitter's verification process, but also does it automatically.

Since the only ground truth available to us is the authenticity of verified users, we conclude that by being able to predict Twitter's verification process, the trust score justifies itself as a good metric for measuring the trustworthiness of users who are not verified.

2.2.2 Identifying Spam Tweets

Having gathered the set of trusted users, we want to find a how the quality users associated with a tweet reflects on the tweet being a spam or not. Users may interact with (and thus be associated to) a tweet in many ways – by clicking the favorite button, by retweeting the tweet or by replying it. Our method uses retweets to determine if the tweet can be trusted or not. More details about the method and the reasons for choosing retweets in the spam detection method are explained in this section.

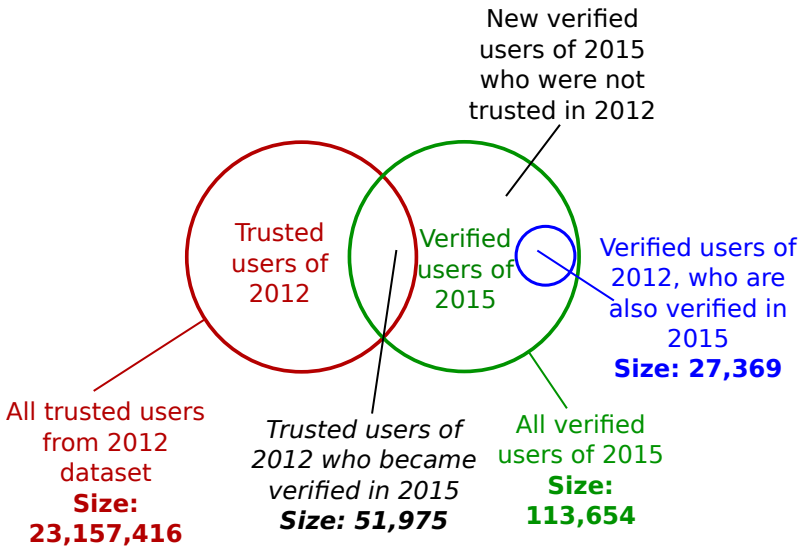


Figure 2.5: *This Venn diagram shows that 60 % of the new verified users of 2015 were from the trusted set of 2012.*

Retweet Chain

We define a retweet chain as the set of "retweeters" for a tweet. A retweeter is someone who retweets a tweet. Thus, for a given tweet, all the people who retweeted the tweet constitutes the retweet chain for that tweet. We will use the notion of retweet chains to deduce if a tweet can be trusted or not (see figure 2.6). But first let us explore the reasons for selecting retweeters as a sign for spam activity.

Significance of Retweets

We defined the three ways users can interact with a tweet in section 1.2.2 – one can retweet a tweet, favorite it or reply to it. Of these three interactions, the most significant is the act of retweeting. Moreover, there are traits in the other two modes of interaction that make them unsuitable for spam detection. The following state the reasons for using retweets (and not the other two) for spam detection:

- **Non-repudiation:** A retweet is a public statement of one's approval of the content of the original content. This means that not only does the user re-distribute the original content to her followers, but the retweet also bears the name of the retweeter. So if a person knows the retweeter to be have a history of sharing interesting tweets, the reader will be more likely to interact with the tweet – and may also retweet it himself. Thus by bearing the name of the retweeter, it is a stamp of approval from the retweeter and thus may influence the reader.
- **No duplication:** By retweeting, the user gives credit to the original content publisher (author of the tweet), unlike copying and pasting (plagiarism) or

Figure 2.6: *A sample tweet showing the number of retweets it has received. The users who created these retweets are called "retweeters" of this tweet and the set of all retweeters for a given tweet is called the "retweet chain" of that tweet. We use the quality of users in the retweet chain to determine if a tweet is spam or not. The intuition is that if a many known legitimate users retweet a given tweet, then the chances of that being a spam tweet is low.*

spamming the same illegitimate tweet from different accounts. Thus retweeting inherently has a non-spam property in its definition.

- **Redistribution of the original content:** Retweeting means redistributing the original content to the retweeter's followers. This is quite different from favoriting, which is more like bookmarking it without engaging other users in doing so. Replies, on the other hand, engage other users but creates new content, which may itself be spam. It is not uncommon to find spam in the replies to popular tweets – that is, spammers try to use the high visibility of popular tweets to spread their spam tweets. Therefore assessing if the original tweet is spam from the replies the tweet gets would not perform well.

- **High visibility:** When a tweet has a long retweet chain, it means that it has a high visibility. That is, because a lot of people retweeted the tweet, all followers of each of those retweeters has received the tweet. This increases the visibility of the tweet drastically. And because it is seen by a lot of people, it would have a significant and possibly detrimental impact (if it is a hoax or rumor, for instance) if it is allowed to spread. Thus, highly retweeted tweets have a strong impact on the society and therefore looking for properties in the retweet chain is a good way to prevent such large-scale manipulation on Twitter and society in general.

- **Affects public opinion:** Activities on Twitter no longer remains in Twitter alone. There are many cases where the real-world is affected by what happens in Twitter. For example, when the non-profit news agency Associated Press' twitter account was hacked and a tweet stating an explosion in the White House was published, the stock market plunged into a drop of 140 points although the news was completely fabricated [41]. This tweet was retweeted more than 1400 times and that fueled the viral spread of this rumor. Because of the many retweets, it affected public opinion, and people started believing it to be true, which in turn caused the stock prices to fall. This is also seen in

many other spheres of society – for example, often the popularity of a product, people's reaction to a movement, or results of an election in a democracy are assessed by analyzing tweets and for each of these cases, among other aspects, highly retweeted tweets play a significant role. Thus, highly retweeted tweets can affect public opinion and trends on Twitter are also affected by retweets. Moreover, unlike favoriting, retweeting is more popular and is used more often in social media analytics and for trends in Twitter.

Spam Detection by Assessing the Quality of Retweet Chains

We assess the "quality" of a tweet's retweet chain to classify it as spam or not. Spam is defined as irrelevant or inappropriate messages sent on the Internet to a large number of recipients [42]. We therefore emphasize on the *quality of a retweet chain* because tweets that are not retweeted much do not have a high visibility, and thus are not that important in the context of spam, which are intended for large number of viewers.

We define the "quality" of a retweet chain as the presence of trusted (or verified) users in it or not. If many trusted (or verified) users retweet a tweet, then it means that the tweet has some interesting information and is most probably authentic (i.e., not spam). This is because of the transitive property of trust itself – trust imparts trust. This is similar to how the acceptance of recommendations in the real world is highly biased by who recommends them. This applies to tweets as well – if we know that trusted users retweet a tweet, it is because in her opinion, that tweet is not spam (because being legitimate, trusted users are expected to not knowingly retweet spam). In this process, we also look at the person who created the tweet in question, and not just the retweeters. From this point, when we refer to the retweet chain, we mean the retweet chain along with the author of the tweet.

There are a few parameters for our method of spam detection. Firstly, we have to choose a threshold trust score t_1 above which we consider users to be trusted (see section 2.2.1). Once that is chosen, then we have to decide a second threshold called "qualifying score for tweets" t_2 to determine if a tweet is trusted – that is we consider a tweet to be spam if there are at least t_2 users in the retweet chain having a trust score of at least t_1. The default values for t_1 and t_2 are both 1, which is also the minimum values for both. We will use these values for the thresholds unless otherwise mentioned in the rest of the report.

2.3 The robustness and on-the-fly nature of the method

Twitter is constantly changing [43], so if our method cannot handle these changes, i.e. if it is not robust, then it will be doomed for failure in the near future. Also, if since tweets on Twitter are getting published in large numbers constantly, if our

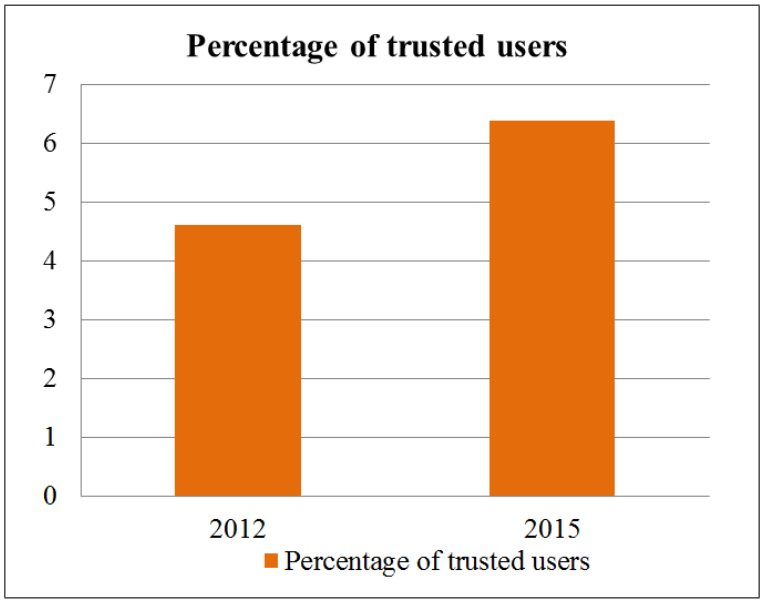


Figure 2.7: *The percentage of trusted users in Twitter has increased from 2012 to 2015.*

method is not real-time and on-the-fly, i.e. if it has to wait for some events to occur and not be able to process tweets as the feed comes in, then our method would not be able to filter spam in time before it spreads to the rest of the social network, which would defeat its purpose. In this section we discuss why our method is both robust and on the fly.

Our method is real-time on the fly because given a tweet, we can find the list of retweets by making another API call to Twitter. More importantly, once we have the set of verified users and the set of trusted users, we can immediately classify a tweet as spam or not with one API call. But there are two issues with this – restricted access to Twitter data and collection of trusted users. The first problem is that using the public API of Twitter we can find at most the 100 most recent retweets for a tweet, and no more. So if a tweet has been retweeted more than 100 times, we will not be able to see the older retweets, but this problem can be mitigated if we have a paid subscription to Twitter's API, and thus it is not really a problem of our methodology. The second problem is that we need the set of trusted users and verified users to be precomputed, and can seem, at first glance, to violate the on-the-fly nature that we wish to achieve. But the trusted set of users changes very slightly over time, because the follow behavior of users does not change drastically. Nevertheless some changes in the trusted set are seen over time (see 2.7). Therefore, what we propose is to periodically collect the trusted set of users so that at any given time, we have a set of fairly accurate set of trusted users which we can use for our spam detection, which makes it on-the-fly. The period of collection of trusted users depends on the type of access to the Twitter API – since the public API is slow, we used the collection of once a week as a baseline. But if a paid subscription is available, it could be more frequent – for example, once a day would be extremely accurate.

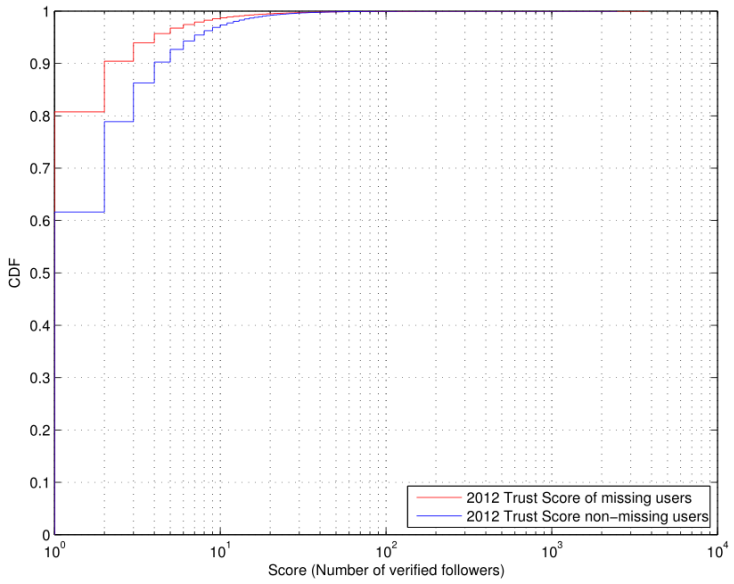


Figure 2.8: *As Twitter changed from 2012 to 2015, there were some changes in the trusted set as well. Some (8.3 %) of the users who were in the trusted set of 2012, were not in the trusted set of 2015 although they had active accounts on Twitter in 2015. This figure shows the CDF of the trust scores of these missing users and the other non-missing users. The users who are missing from the trusted set of 2015 had a lower trust score in 2012, thereby showing that the trust score is a robust metric.*

Our method has to be robust to the changes in Twitter over time and also has to be able to intercept the techniques of spammers so that spammers cannot circumvent our detection method. Following are the reasons why our method is robust to spammers:

- Spammers can create as many accounts as they want, but their accounts will never be verified by Twitter since the verification process of Twitter takes manual background check into account [6].
- It is easy for spammers to publish as many tweets as they want, but they have no control on the follow property of the Twitter social graph. That is, for a spammer it is extremely difficult to change the follow-relationship, which is an integral feature of the Twitter social graph, and that is what our method focuses on. Trusted users are formed from the follow-behavior of the verified users. Since we look for verified and/or trusted users in a retweet chain, a spammer will never be able to circumvent our detection method.

We found that some trusted users of 2012 were not included in the trusted set of 2015, although they exist in 2015. This comprised 8.3 % of trusted users of 2012 and we called them the missing trusted users. This is not an anomaly – instead, this is proof that the trusted set of users is a dynamic, evolving set, and hence our spam detection method is robust to changes in Twitter over time. These missing users may have lost their trusted status because of two probable reasons:

- The verified status of the users who followed the missing user in 2012 were revoked the verified status by Twitter. This may be because these verified users were no longer authentic and genuine as Twitter initially thought. And because of that, the missing user also lost its trust, because trust is seen as a transitive relationship in this project – the verified user trust imparts trust onto the users who she follows. When this verified user loses his credibility of verification, the users followed by him, by the transitive relation, also lose credibility.

- The verified users who followed the missing user stopped following him/her. This maybe because the verified users realised that this missing user is in fact an imposter/fraud, and were thus they were not interested in them anymore. This is a strong validation of the dynamic nature of the trusted set.

We found that in 2012, the missing trusted users had a lower trust score than the non-missing trusted users, as shown in figure 2.8. The figure is a CDF plot of the 2012 trust score of the missing users and the non-missing users. It shows that the users who are missing from the trusted set of 2015 had a lower trust score in 2012. Thus, trusted users having lower trust scores can become untrusted over time more easily than trusted users having higher trust score, and this is exactly the semantics behind the trust score – higher the trust score, more trusted the user is, and thus less likely to become untrusted over time. Since the trust score is robust, our spam detection method is also robust.

Chapter 3

Testing the Method

Now that we have discussed our method, in this chapter we will show how we evaluated our method using different methods and datasets to correlate our spam detection method with existing methods or, if a correlation was not found, to explain the reason for the absence of the correlation. We have to also show that our method is robust and can be applied on the Twitter feed in real-time (on-the-fly). To achieve these two goals, we have tried a number of techniques which are discussed in this chapter.

3.1 Finding a Definitive Test for Spam Detection

Before we go into the discussion of the methods we used for testing our spam detection algorithm, it is important to point out that testing a spam detection method is very difficult. We did not have a test set of tweets in which tweets were already definitively classified as spam or not. This is more than just unavailability of a classification dataset – the root of the problem is that there is no spam detection method that is 100 % accurate. Because of that, since we are developing a new spam detection method, we should be hopeful that our method detects some unique properties of spam that other methods cannot detect, which by definition makes it impossible to have a proper test set. Manual verification could work, but we are working with the scale of millions of tweets, and manual verification would not be viable, though it would work for a subset (as we show later). Therefore we had to use alternative methods based on the *properties* of spam, rather than a binary classification mechanism, because no classification is 100 % accurate. With that in perspective, we now discuss some of the techniques we explored.

3.2 Preliminary Tests

In this section we describe our first attempts at testing our method. These testing methods were not conclusive in justifying our spam detection method, but it is interesting to discuss them because they bring new information about our method, and as discussed in the previous section, even failed tests bring new information to light – that is because no spam detection method is 100 % accurate.

3.2.1 Using the Number of Retweets to Detect Spam

The first technique to test our method of spam detection was to see if it correlates to the number of retweets tweets have. The idea behind this is that since we are testing the quality of retweets to decide if a tweet is spam or not, if most tweets do not have enough retweets, it could potentially be classified as spam – in which case it would be a wrong classification. Moreover, even if the classification was correct, if the number of retweets was correlated to our method, then we could as well use the number of retweets instead of using the trust score. Therefore we first had to show that the number of retweets was not a good method for spam detection.

We sampled a set of 137,896 tweets from the dataset obtained from Vigiglobe [45]. We chose tweets which had less than 100 retweets. Because the Twitter API returns only a list of latest 100 retweets, this ensured that we can study the tweets' full retweet chains. We then queried the Twitter API and obtained the full retweet chains for these tweets. As explained before, by "retweet chain" of a tweet, we mean all the users who either retweeted the tweet or created the tweet. Upon analyzing the retweet chains, we found that out of 137,896 tweets, 119,627 had at least one trusted (or verified) user in their retweet chains. Thus, around 2.1 % of the tweets were completely untrusted or spam. We then plotted the CDF of the number of retweets for each retweet chain, based on if the tweets had a trusted (or verified) user in their retweet chain or not. Figure 3.1 shows that there are a few spam tweets which are retweeted a lot. On further inspection of this unexpected behavior, we found that:

- Some tweets could not be found (using Google search, Twitter web search, Twitter Search API): They were either deleted by the creator or removed by Twitter, due to some reason.
- Some tweets had the exact same content: This further indicates spam activity – spammers often create multiple acoounts and publish the same content as individual tweets to get maximum visibility. Most of these tweets had a small retweet chain, and because they are individial tweets, they were treated by this testing method as individual tweets. This decreases the efficiency of the testing method. It might seem at this stage that our spam detection method would also suffer from this problem in the same way, but it does not, as we will describe later in this chapter (see section 3.3).

- Some tweets were made in a reply to a very popular tweet: Because popular tweets have a high visibility, we observed that spammers replied to popular tweets to get noticed. By virtue of the popularity of the original tweet, they were retweeted more than expected but since this testing method only looks at the number of retweets and not the quality of the retweets, such tweets would slip through and be classified as non-spam.
- Some tweets mentioned a famous (and often verified) Twitter account: Similar to the reason above, we found that spam tweets often mention a famous celebrity to get noticed, and hence scavenge some retweets. This testing method would also fail to detect such spamming techniques.
- Some tweets contained a popular hash tag: Similar to mentioning a famous Twitter account, mentioning trending hashtags causes an increased amount of retweets, which would slip through if the number of retweets was used as a deciding factor.

Because the number of retweets can be influenced by such tricks, we concluded that the number of retweets is not a good measure of estimating spam behavior. Moreover, we queried the Twitter API to find the users who retweeted the highly retweeted untrusted tweets (tweets having at least 20 retweets). We found that there were 722 such users exist in our collection, but when we queried Twitter we found that only 700 (97 %) of them exist. In contrast, when we found the number of users who retweeted the highly retweeted trusted tweets were 493,815, but only 492,556 (99.75 %) were found to exist when we queried the Twitter API. That is, although the number of retweets may be high, the users retweeting them may be suspicious, which is why Twitter may have deactivated their accounts. In conclusion, we cannot use the *number* of retweets as an indication of spam activity – the *quality* of retweets is expected to be a much better indication. Thus we need to look for other testing methods.

3.2.2 Detecting Spam from Periodic Tweets

On Twitter, there are many active accounts which are run by computer programs and are called "Twitterbots" [46]. We manually found Twitterbots that publish tweets periodically, for example, every five minutes. These tweets, we found on manual inspection, are spam and have no information content. We tried to find these Twitter bots that automatically publish tweets periodically. We wanted to find periodic bursts of tweets, which might correspond to spam bots waking up. We argued that Twitterbots that had the restriction of publishing something every period, will soon run out of useful tweets and will resort to publishing tweets that would classify as spam. We therefore tried to use the period of publication of tweets to find these accounts so that we can then use them to evaluate our spam detection method. We tried to find a periodicity in the number of tweets published every millisecond which we hoped would be the result of all Twitterbots waking up. But unfortunately there were no discernible patterns in the times of publishing of tweets,

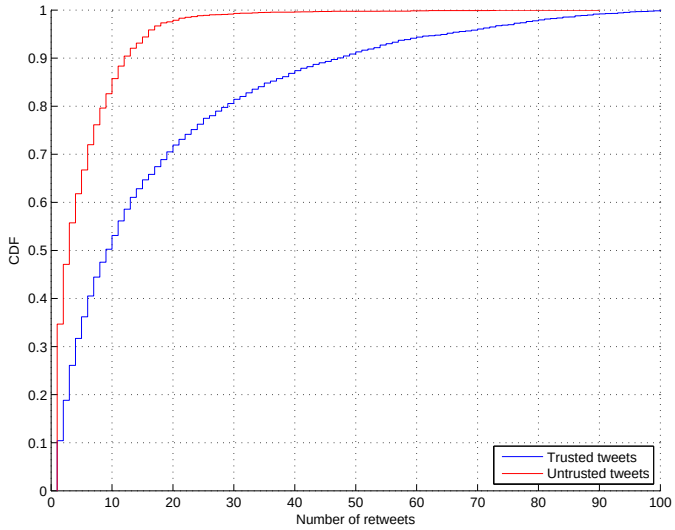


Figure 3.1: *This figure shows the CDF of the number of retweets for trusted tweets and untrusted tweets. We can see that there are a few untrusted tweets which are retweeted a lot.*

as seen in figure 3.2. The problem was with the assumption that all Twitterbots have the same period, which would give rise to the burst of tweets being published.

3.2.3 Detecting Spam using Trigger Words

The most common and most used method for spam detection is keyword filtering – spammers often use words that are either offensive, out of context, or simply blatant advertisement. For example, `#followback`, `free download`, `fuck`, etc. are a few common words often found in spam tweets. Since this technique is used widely, we wanted to correlate this method with ours and see how well our method performs compared to it. We thus tried to find spam activity by looking for keywords that were in no way related to the theme of our dataset of tweets "Microsoft" or were derogatory in some sense. We called these words "trigger words".

We wanted to find the most frequently occurring trigger words in our dataset. Since spammers tweet in large numbers, trigger words which occurred frequently would be a good indication of spam activity. Thus we generated a word cloud for all the words occurring in our tweet dataset (see figure 3.3), in which the font size is proportional to the frequency of occurrence in the dataset. From the most commonly occurring words in our dataset, we selected the ones that were not related to the topic of the dataset "Microsoft", and were seen very frequently. Once we had the trigger words, we found the tweets containing these words, and expected them to be spam. We tried to compare these tweets with the results obtained from our spam detection method. But the results we found upon comparison were shocking. We found that there was a higher percentage of non-spam tweets containing these trigger words than spam tweets (see table 3.1). That is, a higher percentage of tweets which our method declared as non-spam contained the trigger words used for spam detection by most

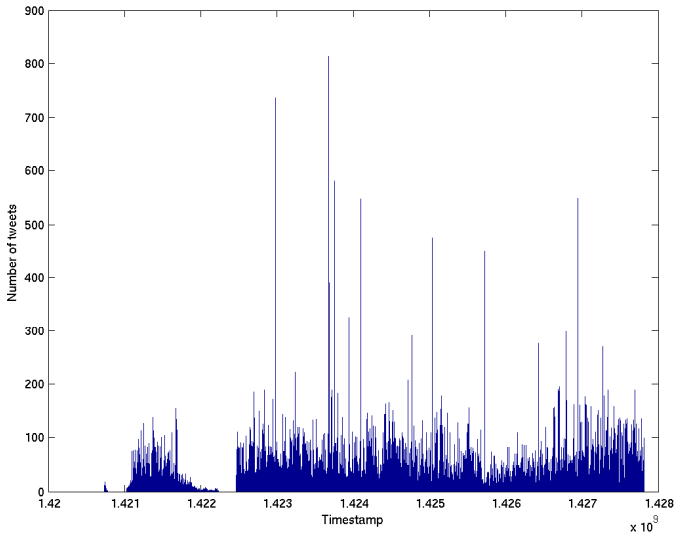


Figure 3.2: *This figure shows the number of tweets by time (in seconds). In X axis are the timestamps and in Y axis, the number of tweets in one second.*

social media analytics analyses. This meant that either our method was missing some property of spam tweets, or that our method could detect some property of spam tweets which was hidden to the spam detection method of using trigger words.

As seen from table 3.1, a higher percentage of non-spam tweets (from our method) contain trigger words than those declared as spam tweets by our method, and this was counter-intuitive. Therefore we looked further into this anomaly and we found that using trigger words as a spam detection method often wrongly classifies tweets and therefore is not a good way to find spam. This is because although some keywords may seem derogatory on their own, the entire text of the tweet may not be illegitimate – and this is only understandable when we look at the rest of the tweet, its grammar and its sense – that is, a complete semantic analysis of the tweet, which is difficult to do accurately for a computer. We found instances of such tweets as shown in figure 3.4 in which we show a tweet that contains a trigger word "`porn`" and a trigger hashtag "`#WTF`" but is not spam because its content is legitimate and it talks about the real event when .porn domain names for websites were being sold. Such tweets would be classified as spam by methods that only rely on the presence of trigger words, but out spam detection method did not mis-classify it because the author of the tweet is a trusted user and so the tweet was not classified as spam by our method. It should be noted here that the failure of this method casts a shadow on the efficiency of social media analytics that rely on studying trigger words not only for spam. This is the case for many companies in the industry today, including Vigiglobe [45], which is one of the eighteen companies in the world to have direct unrestricted access to Twitter's data.

Figure 3.3: *This figure shows the top 1,600 most frequently occurring words in our dataset of tweets related to the topic "Microsoft" (font size is proportional to frequency). Most of the words are related to "Microsoft" but there are many unrelated and out-of-context words like "shoes", "fashion" and "leather".*

Table 3.1: *Percentage (rounded to the nearest 10^{-2}th place) of non-spam and spam tweets (as detected by our spam detection method) that contain trigger words.*

Trigger word	*Percentage of* ***spam tweets*** *containing the trigger word*	*Percentage of* ***non-spam tweets*** *containing the trigger word*
bullshit	0.01 %	0.02 %
hack	0.17 %	0.36 %
drunk	0.01 %	0.02 %
torrent	0.22 %	0.02 %
hack	0.17 %	0.31 %
followback	0.02 %	0.03 %
sweetheart	0.00 %	0.00 %
webcam	0.02 %	0.04 %
giveaways	0.02 %	0.07 %
fuck	0.25 %	0.67 %
cheap	0.11 %	0.14 %
bomb	0.02 %	0.03 %
gay	0.06 %	0.07 %
lovers	0.01 %	0.02 %
sale	0.23 %	0.33 %
firmware	0.03 %	0.03 %
mom	0.16 %	0.50 %

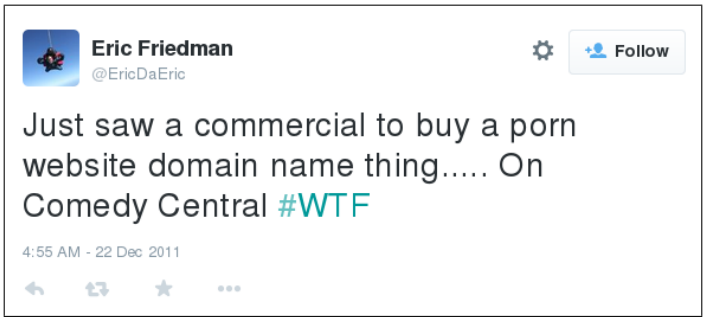


Figure 3.4: *This tweet, published by a Twitter user named Eric Friedman, contains a trigger word "porn" and a trigger hastag "#WTF", but is not a spam tweet, as it talks about the legitimate incident when the .porn domain name started being sold. Spam detection methods that use the presence of trigger words would classify this legitimate tweet as spam, but our spam detection method did not, because Eric Friedman is a trusted user. This shows that keyword analysis to find spam on Twitter is not a good spam detection method although it is used widely in the industry to detect spam.*

3.3 Using Content Duplication for Testing

The testing methods we discussed till now have not been able to show that our spam detection method works. Each of the aforementioned testing schemes were, in some way, unsuitable for verifying our spam detection method. It is useful to mention here, once again, that in this chapter we are discussing ways to show that our spam detection works well. The goal of this entire work is to find a spam detection method that is robust to the changing parameters of Twitter over time, robust to the spamming techniques used and can be applied on-the-fly. As we explained in section 2.3, our spam detection method is both robust and can be applied on-the-fly. In this section, we will show how our method correlates to a different method that we devised to detect spam activity. This new testing method is neither robust nor can it be applied on the fly and therefore it is not suitable for spam detection in Twitter – we simply use it to correlate our results with this technique. We applied it statically on our dataset to obtain the result solely to be able to measure how well our spam detection method performs.

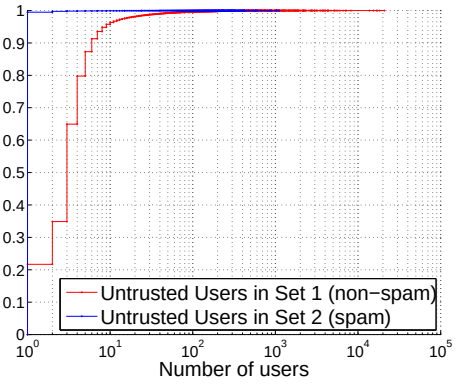
Since using trigger words did not work as a testing method (as discussed in section 3.2.3), we tried to analyze tweets as a whole – by using the entire content of tweets. Many spam tweets often have the exact same content. This is because spammers often create multiple accounts and publish the same tweet from all these accounts in the hope of getting a higher visibility. As we described in section 2.2.2, this kind of behavior is exactly what makes retweets a good choice to detect non-spam behavior and thus the goal here is to use this known spam property and see if our method can detect that. In order to do this, we compared the content of each tweet in our dataset with every other tweet in the dataset and counted the number of duplicates it had. In the scope of analyzing the content of tweets, there have been a lot of work on semantic analyses of tweets and machine learning [44, 47]. Such semantic analysis of tweet content is not perfect and is out of scope for this work.

For each tweet in our dataset, we counted the number of times the exact same content occurs in our dataset. We can have two cases where the content of two tweets are exactly same – (i) the duplicate tweet is a retweet of the same original tweet or, (ii) the duplicate and original tweets may have been published by the same spammer from different accounts. The second case (termed "content duplication") is spam in most cases. The reason for content duplication can be any of the following:

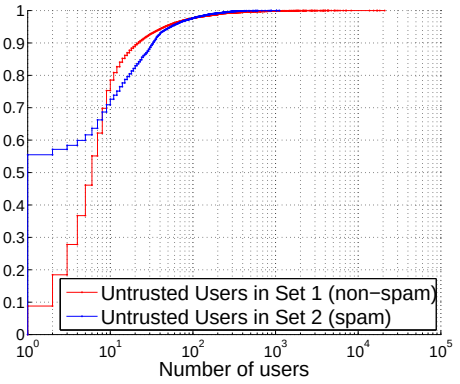
- To get high visibility, without providing any useful content: The idea of spam is to reach a large audience without any credible information. This is similar to the quantity vs quality issue – when there is no quality, one wants to market a product in large quantities to prevent a loss.
- The spammer may try to influence multiple components of the Twitter social graph that are not directly linked to get maximum visibility. The spammer can do this by starting a spam campaign at different communities/groups on Twitter, by creating fake accounts (which is also a reason we focus on users who are trusted – after all, users create tweets on Twitter).
- Groups of spammers may try to help each other out by retweeting/copying each other's spam tweets in a mutually beneficial relationship.

To see if our spam detection method can detect such content duplication, we divided our dataset of tweets into two categories for the spam analysis – Set 1 and Set 2. Set 1 consists of tweets which have more retweets than duplicates. Set 2 consists of tweets which have no retweet at all. On top of this, we used a threshold t_3 for creating the sets. Only tweets with at least t_3 occurrences (duplicates or retweets) are kept in the two sets. For example, if $t_3 = 5$, in Set 1, we will have only tweets whose $num_{retweets} + num_{duplicates} >= 5$ and in Set 2, we will have tweets whose $num_{duplicates} >= 5$, since by definition, tweets in Set 2 have not been retweeted. Here, $num_{retweetes}$ means the number of retweets of the tweet and $num_{duplicates}$ means the number of times the content of the tweet has been duplicated as another tweet.

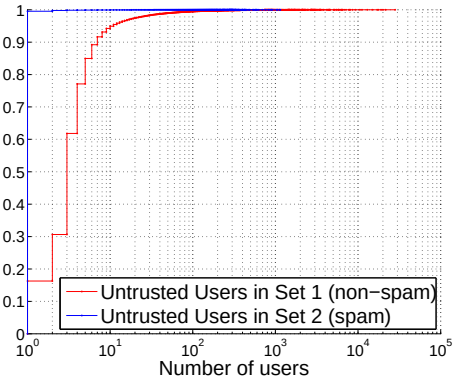
We expect that Set 1 would be less susceptible to spam activity than Set 2 because Set 2 contains only tweets which have not been retweeted but have been duplicated at least t_3 times) – which, as we discussed before, is spam activity. Figure 3.5 shows six CDF plots each of which corresponds to a trust score (t_1) and a threshold (t_3). We are interested in the number of trusted and untrusted users associated with tweets in the two sets (association with a tweet means the act of either retweeting or publishing the tweet). Since we know that Set 2 exhibits spam behavior, if we find that there are more untrusted users in Set 2, then we can claim that our method can identify such spam behavior. Therefore, in each of the six plots in figure 3.5, we show the CDF of the number of untrusted users in each of the two sets. What we want to show here is that for a given value of trust score (t_1), as we increase the threshold (t_3) from 1 to 8, the number of untrusted users in Set 2 increases very drastically compared to Set 1. Thus, as the number of duplicates increases, the number of untrusted users in Set 2 also increases faster than it does for Set 1. Therefore, we can say that our method can detect spam activity.



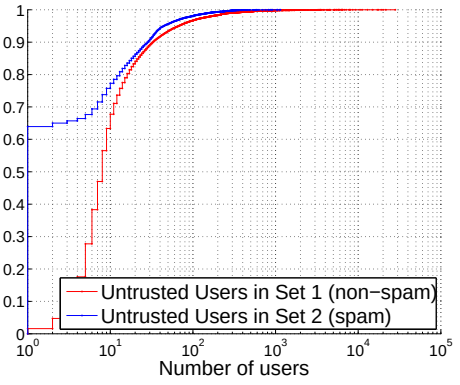
(3.5.1) $t_1 = 1, t_3 = 1$



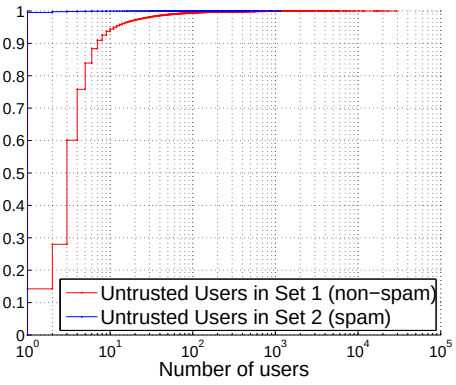
(3.5.2) $t_1 = 1, t_3 = 8$



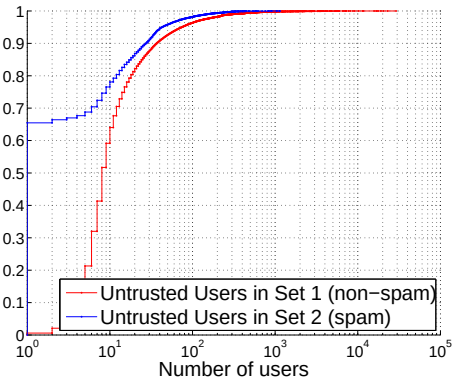
(3.5.3) $t_1 = 5, t_3 = 1$



(3.5.4) $t_1 = 5, t_3 = 8$



(3.5.5) $t_1 = 10, t_3 = 1$



(3.5.6) $t_1 = 10, t_3 = 8$

Figure 3.5: *Each figure is the CDF of the number of untrusted users associated with tweets in Set 1 and Set 2 for a given trust score (t_1) and a threshold (t_3). Set 1 contains only tweets which are retweeted more than duplicated, and Set 2 contains tweets which have not been retweeted at all. The threshold t_3 is used such that only tweets that have at least t_3 occurrences (duplicates or retweets) are included in the sets. For every trust score, as the threshold increases from 1 to 8, we see that the number of untrusted users associated with tweets of Set 2 (spam set) increases whereas that for Set 1 (non-spam set) changes only slightly, thereby showing that the Set 2 (spam set) has more untrusted users than in Set 1 (non-spam set). We showed here that spam properties like duplication of tweets can be detected by our method.*

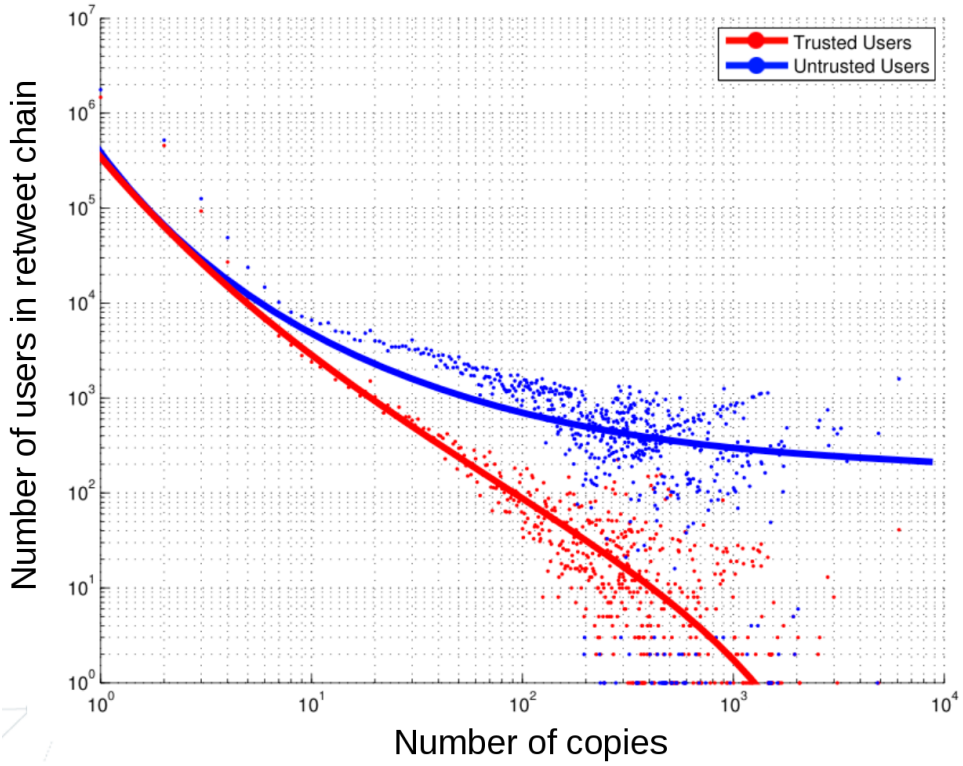


Figure 3.6: *This figure shows that with the increase in the number of copies (duplicates) of a tweet, the number of trusted users in retweet chains decreases to zero but the number of untrusted users remains fairly constant after a point. This shows that our method can detect spam behavior on Twitter.*

To provide further support in favor of our method, we computed how the number of trusted users and untrusted users in the retweet chain changes as the number of duplicates of tweets increases. As we see in figure 3.6, with the increase in the number of duplicates, the number of trusted users associated with tweets decreases to zero, but the number of untrusted users does not. This shows further that with content duplication, which is seen to be spam behavior in most cases, the number of trusted users associated with duplicated tweets decreases further and further with the increase in the number of duplicates. Thus, this is another validation that our method can detect spam properties in Twitter.

Chapter 4

Conclusion

We have now showed that our method can detect spam activity on Twitter. In this chapter, we discuss the possibilities of the next stages for this work.

Future Work

For further confirmation, we also want to compare the results of our method with another method currently being developed by other members of our team under my supervisor Dr Arnaud Legout. This method, called "The Hashtag Graph", tries to study relationships between hashtags and keywords in tweets to detect anomalous behavior that point to spam. It will be interesting to compare our results with the results of this method because this method approaches the problem of spam in a different way than ours – i.e., by looking for critical behavioral traits in way hashtags are used in Twitter. If a strong correlation is seen between two methods that are completely unrelated in their methodology, it would be a very strong result for both methods.

We are also interested in building an online spam-detection service that would take the URL of a tweet and categorize it as spam or not. This would help the end user be more aware on Twitter. It could also be used to take users' feedback to our classification – if our method makes a mistake in the classification, it could be reflected by the user in the feedback, which we can use as a crowd sourcing mechanism to fine-tune our method.

As with most spam detection mechanisms used today, our method can be tuned to the perfect balance between having too many false positives or too many false negatives. We use two parameters, namely a threshold trust score for users (t_1) and a qualifying score for tweets (t_2). As described in section 2.2.2, the threshold trust score is the minimum score that users should have to be included in the set of trusted users and the qualifying score for tweets is the minimum number of trusted users in the retweet chain required for a tweet to be considered as legitimate and not spam.

The tuned performance using these two parameters will offer some flexibility in the spam detection process as required by the real world situation. Such real-world scenarios and thereby our method's performance in the wild is yet to be evaluated.

Most spam detection techniques for Twitter take a tweet-centric approach, that is, they try to detect is a tweet is spam or not based on properties of the tweet like the presence of some keywords, semantic analysis of the content, types of hashtags present etc. Our method takes a user-oriented approach in spam detection by first profiling users. Then, based on the quality of users who have associated themselves with a given tweet, we perform our classification of the tweet as spam or not. The intuition behind this approach is that since tweets are ultimately published by users, the quality of the tweet should be apparent from the quality of the users interacting with a tweet. A spammer will always try to publish spam tweets and a legitimate user will always try to retweet what interests him. In this work we have therefore viewed tweets as interactions between users and the quality of these users is used to assess if the tweet is spam or not.

Conclusion

Over the years, as Twitter grew in size and popularity, spam activity on Twitter also kept increasing. Because Twitter has a high influence on the society today, it is very important to keep Twitter as free from spam and illegitimate content as possible. Twitter has not been able to solve the problem of spam perfectly, which is the reason trends on Twitter are susceptible to manipulation and misinterpretation which can be very harmful to the society as a whole. Thus it is very important to keep Twitter as free from spam and illegitimate content as possible. But keeping Twitter completely free from spam has not been successful even for Twitter themselves. Thus, there is a need for developing methods like ours that will handle spam filtering on Twitter from a different perspective (by assessing the quality of users as in our case). Without proper spam detection, the high influence of Twitter on the society could result in a large scale manipulation of users or misinterpretation of opinion of users on Twitter. This can have a severe effect like the many hoaxes, stock market crashes and other fraudulent activities that were fueled by Twitter. Our method of spam detection aims to detect and filter spam before it reaches the masses to create havoc like these.

Bibliography

[1] Wikipedia article on Twitter. http://en.wikipedia.org/wiki/Twitter.

[2] About Twitter. https://about.twitter.com/company.

[3] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? WWW'10, Raleigh, NC, USA, May 2010.

[4] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph. ACM SIGMETRICS'14, June 16-20, 2014, Austin, Texas, USA.

[5] Mashable.com. Eight social media hoaxes you fell for this year. http://mashable.com/2012/11/05/social-media-hoaxes

[6] FAQs about verified accounts. https://support.twitter.com/articles/119135-faqs-about-verified-accounts.

[7] Mashable.com: Facebook Is Most Popular Social Network for All Ages; LinkedIn Is Second [STUDY]. http://mashable.com/2011/11/04/facebook-most-popular-forrester

[8] Maksym Gabielkov, Arnaud Legout. The Complete Picture of the Twitter Social Graph. ACM CoNEXT 2012 Student Workshop, Dec 2012, Nice, France.

[9] N. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, K. P. Gummadi. Inferring Who-is-Who in the Twitter Social Network. 4th ACM SIGCOMM Workshop On Social Networks (WOSN), Helsinki, Finland, August 2012.

[10] TechCrunch.com. Twitter picked up 16M active users in Q2. http://techcrunch.com/2014/07/29/twitter-q2-user-growth

[11] Twitter help center. Using Twitter lists. https://support.twitter.com/articles/76460-using-twitter-lists#

[12] Michael Nielsen. How to crawl a quarter billion webpages in 40 hours. http://www.michaelnielsen.org/ddi/how-to-crawl-a-quarter-billion-webpages-in-40-hours

[13] H. Liu, E.-P. Lim, H. W. Lauw, M.-T. Le, A. Sun, J. Srivastava, Y. A. Kim. Predicting trusts among users of online communities: an epinions case study. ACM Conference on Electronic Commerce (EC2008), Chicago, 2008.

[14] Y. Matsuo and H. Yamamoto. Community gravity: measuring bidirectional effects by trust and rating on online social networks. WWW, 2009.

[15] M. Richardson and P. Domingos. Mining knowledge sharing sites for viral marketing. In 8th ACM SIGKDD, 2002.

[16] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In ACM WSDM, 2010.

[17] D. Ramage, S. Dumais, and D. Liebling. Characterizing Microblogs with Topic Models. AAAI Conference on Weblogs and Social Media (ICWSM), 2010.

[18] About Twitter's suggestions for who to follow. `https://support.twitter.com/articles/227220`

[19] D. Kim, Y. Jo, I.-C. Moon, and A. Oh. Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users. ACM CHI Workshop on Microblogging, 2010.

[20] A. Pal and S. Counts. Identifying topical authorities in microblogs. ACM Conference on Web Search and Data Mining, 2011.

[21] Alina Quereilhac, Mathieu Lacage, Claudio Freire, Thierry Turletti, Walid Dabbous. NEPI: An Integration Framework for Network Experimentation. Software, Telecommunications and Computer Networks (SoftCOM), 2011.

[22] Simi John. Morgan Freeman Death Hoax: Bruce Almighty Star Denies Death Rumours. International Business Times, Oct 2012.

[23] Mashable.com. Police: Teenage Girl's Viral Tweet Was Kidnapping Hoax. `http://mashable.com/2012/10/01/teenage-girl-tweet-kidnapping`.

[24] Twitter.com. Rate limits: Chart. `https://dev.twitter.com/rest/public/rate-limits`.

[25] Huffington Post: Eleazar David Melendez. Twitter stock market hoax draws attention of regulators. `http://www.huffingtonpost.com/2013/02/01/twitter-stock-market-hoax_n_2601753.html`.

[26] Tweet published by Twitter France on the `#JeSuisCharlie` hashtag. `https://twitter.com/twitterfrance/status/552966270866706434`.

[27] Twitter's definition of "following" and "followers". `https://support.twitter.com/articles/14019#`.

[28] Spreading of the `#JeSuisCharlie` hashtag all over the world. `http://srogers.cartodb.com/viz/123be814-96bb-11e4-aec1-0e9d821ea90d/embed_map`.

[29] Mashable: History of Twitter. `http://mashable.com/2011/05/05/history-of-twitter`.

[30] Wikipedia article on the *Je suis Charlie* campaign. https://en.wikipedia.org/wiki/Je_suis_Charlie.

[31] Sergey Brin, Lawrence Page. The anatomy of a large-scale hypertextual web search engine. WWW7 Proceedings of the seventh international conference on World Wide Web 7, Amsterdam, The Netherlands 1998.

[32] Wikipedia article on microblogging. https://en.wikipedia.org/wiki/Microblogging.

[33] Wikipedia article on hashtags. https://en.wikipedia.org/wiki/Hashtag

[34] Twitter blog article on "Best practices for journalists". https://blog.twitter.com/2012/best-practices-for-journalists.

[35] Twitter blog article on "To Trend or Not to Trend". https://blog.twitter.com/2010/trend-or-not-trend.

[36] FAQs about trends on Twitter. https://support.twitter.com/articles/101125#.

[37] Forbes article on "The most notorious fake Twitter accounts". http://www.forbes.com/2010/08/02/bp-angelina-jolie-technology-twitter.html.

[38] Twitter's guide to reporting spam on Twitter https://support.twitter.com/articles/64986#

[39] Twitter's definition of retweets. https://support.twitter.com/articles/77606#

[40] Wall Street Journal's article on "Identifying spam is tricky for Twitter " http://www.wsj.com/articles/SB10001424052970203686204577114613630000908

[41] CNN news article "False White House explosion tweet rattles market". http://buzz.money.cnn.com/2013/04/23/ap-tweet-fake-white-house/?iid=EL.

[42] Dictionary definition of "spam". http://www.merriam-webster.com/dictionary/spam.

[43] Yabing Liu, Chloe Kliman-Silver, Alan Mislove. The tweets they are a-changin': Evolution of Twitter users and behavior. Association for the advancement of artificial intelligence, 2014. The eighth international AAAI conference on weblogs and social media.

[44] Juan Martinez-Romo, Lourdes Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. Expert systems with applications Journal.

[45] Vigiglobe – one of the 18 startups worldwide to have direct access to Twitter's data. http://vigiglobe.com/v2/.

[46] Wikipedia article on Twitterbots. `https://en.wikipedia.org/wiki/Twitterbot`.

[47] Zi Chu, Indra Widjaja, Haining Wang. Detecting Social Spam Campaigns on Twitter. International Conference on Applied Cryptography and Network Security (ACNS), 2012.