An Analysis on The Network Structure of Influential Communities in Twitter

by

Adam Schunk

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Masters of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2019

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Over the past years online social networks have become a major target for marketing strategies, generating a need for methods to efficiently spread information through these networks. Close knit communities have developed on these platforms through groups of users connecting with like minded individuals. In this thesis we use data pulled from Twitter's API and from simulations designed to mirror the Twitter network to pursue an in depth analysis of the network structure and influence of these communities. Through this analysis we draw several conclusions. First, the influence of users in these communities is correlated to the total number of followers in their neighborhood. Second, influential communities tend to be more tightly clustered than other areas of the network. Using these observations, we develop an algorithm to detect influential communities in Twitter and show that correctly prioritizing connections yields significant gains in message visibility.

Acknowledgments

I would like to thank Professor Kate Larson, PhD, for the opportunity to work with and learn from her.

To my family, friends and teachers: I thank you for providing me with support and knowledge throughout my life and for helping me when I go astray.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Efficient spread of information in social networks is a crucial component to many areas of industry and has been the subject of much interest in the past decade [6]. As social networks such as Facebook and Twitter increased in popularity they created platforms for entities to spread their messages by reaching many users. Research, driven by a desire to understand the factors that effect the spread of messages and how social networks can be leveraged to influence consumers, has proven invaluable in marketing [22], executing information campaigns [20], and cultivating relationships with consumers [37]. A significant portion of this research is directed towards identifying influential users and their characteristics.

Extensive research exists detailing different measures of user influence in social networks graphs [63, 59]. Two of the most frequently used measures in the analysis of traditional social graphs are *degree* and *closeness*. *Degree* represents the number of direct connections a vertex has while *closeness* measures the sum of the shortest paths from a vertex to the rest of the graph. Other tools such as *Eigenvector Centrality* [12] and *betweenness* [25, 15] are effective measurements to understand the structure of networks. Individually, these methods are applicable to measuring the structure of online social networks but fail to capture many of their characteristics.

Past work in Twitter focused on topics such as determining key influential users [17], optimal content for spreading messages [16], and user motivation for increasing message visibility [45]. While these methods show success [66, 8] it is acknowledged that they do not fully encompass the complexities of the topic [57]. We believe that a component missing

from the current literature is an applied analysis on the effect Twitter's network structure has on the influence that communities exert on the network around them.

In this thesis we combine both real world and simulation data to identify correlations between the properties of well clustered groups of users and the influence they exert on message propagation. Through our investigation we have developed an algorithm that can be used to predict the relative influence of groups within a social network and provide recommendations on what areas of the network to target in order to spread information efficiently.

1.1 Twitter

Within the past decade Twitter has become one one the most prominent online social networking services. Twitter allows users to publish short 140 character messages, called tweets. In addition to text, tweets can include *hashtags* to mark them as belonging to a specific topic and *mentions* to tag specific users. Users can elect to subscribe to other users to receive updates from them such as the messages they publish. The Twitter network can be described as a directed graph where vertices represent individual users and the edges between them their relationships. An edge is formed when one user chooses to subscribe to another, and information can only flow through the graph by following the direction of each edge. If user A subscribes to user B then A becomes a follower of B and B a followee of A, allowing information to flow from B to A. For the purposes of this paper we define the *in-degree* and *out-degree* of each user to represent the number of people they follow and their followers respectively. The *mutual degree* of a node is the number of users they follow who also follow them back. This type of relationship allows information to flow bidirectionally between users.

A user's homepage displays messages published, in real time, from everyone they are subscribed to. Users have the choice to *comment* on, *like*, or *retweet* any tweet displayed on their homepage. All of these actions will send a notification to their followers, so each action contributes to the propagation of the message in varying degrees. *Comments* and *likes* allow users to reply to a message or indicate support for it, but they are not the primary force behind message propagation in Twitter. When a user retweets a message, it is sent to all of their followers who then also have the ability to perform the same three actions. Retweeting differs from both other actions because its sole purpose is to share

a message. Even though *likes* and *comments* spread the message to a user's followers, retweeting has become the primary method of message propagation within Twitter.

Prior to the development of Twitter's retweet system users would copy and paste tweets they wanted to share and tag them with RT, for retweet, or add a *mention* to the original author. This caused issues when attempting to retweet messages that were at, or close to, the 140 character limit allowed by Twitter. Due to this size restriction large tweets needed to be altered by removing content so that the tags would fit. In some cases users would simply not retweet anything that would require editing [14]. After the official adoption of retweeting users were able to share any tweet they found interesting with the simple click of a button regardless of the message size.

Even small numbers of retweets can have a large effect on tweet visibility. Kwak, et al. concluded messages retweeted even a single time reach on average 1000 users, regardless of the number of followers of the initial author [34]. Twitter users have on average 20 followers [44], indicating that the Twitter network has naturally evolved to a state where retweeting has become the predominant method for information dissemination. Since commenting is less significant in message propagation and it is not possible to obtain information on who *liked* a tweet, focusing on retweets is the most reliable method to measure how widely visible a message has become. Researchers measure influence of users by how likely they are to be retweeted and use this information to define relationships between characteristics of users, their tweets, and the influence they exert [16, 66, 4].

1.2 Motivation for Twitter Analysis

With almost 350 million users and 500 million tweets sent daily the Twitter network is one of the the largest and most active online social networks in existence [62, 54]. This scope makes Twitter an attractive opportunity for entities that benefit from spreading information quickly and widely. For example, many companies leverage Twitter to directly connect them to their target consumers for advertising and increasing visibility. Market need has driven research to evaluate which factors contribute to influence within Twitter and how these factors can be quantified.

Twitter has gained significant traction in the commercial advertising business [49, 19], despite many individuals reporting feelings of disinterest and annoyance in response to

online advertisements [51]. According to Twitter's 2017 year end report they earned more than $2 billion in advertising revenue [61] in that year alone. A portion of this revenue was obtained from companies using their own Twitter accounts to forge relationships with their consumers in an effort to influence them on a personal level. When a user follows a company they demonstrate a loyalty or fondness for the products that they subscribed to. In turn, this loyalty typically makes them more receptive to advertisements and more open to spreading positive messages [6]. These methods have proven effective in generating exposure [55] which has encouraged researchers to focus on developing more effective means of social media advertising by striving to understand the causes of influence within a network and identify features common among tweets with the highest visibility.

In addition to commercial uses many other groups and organizations have demonstrated an active interest in spreading messages via Twitter. Politically motivated messaging is a highly visible example. While much of the information on Twitter is accurate, the social media platform been used for disinformation campaigns. The rampant spread of misinformation during the United States 2016 election, often dubbed "fake news", was also a motivating factor for selecting this thesis topic [3]. The fact that communities have a significant impact on message propagation creates the potential for false narratives to spread through influential communities and infect other parts of the network. We wished to understand the methods by which these communities influence the surrounding network and their role in the spread of "fake news".

1.3 Related Literature

1.3.1 Influence

When determining influence within a network it is vital to understand the method by which information flows through that network. There are several categorizations of information dispersal, one of the most prominent is the one-step vs two-step classification [36]. The one-step method involves messages being directly broadcast to the intended audience. Media such as a newspaper or televised news programs feature a one-step system as they reach a wide audience and once seen are not generally propagated much further. The one-step method works well to reach a large base but suffers when trying to expand past the initial

step. These types of media are mainly consumed by those who already believe in the message being spread [46, 56].

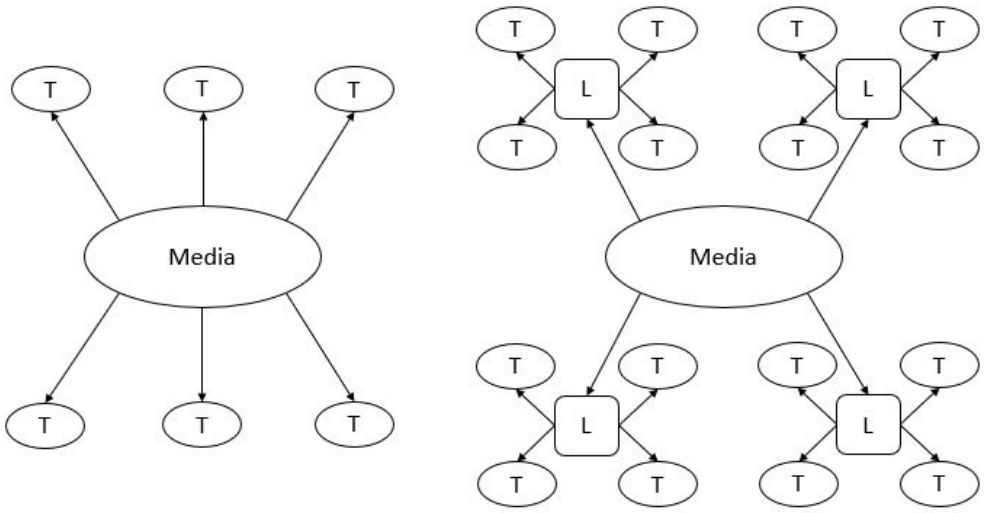


Figure 1.1: One-Step versus Two-Step method. T represents members of the target audience, L represents a community leader.

The two-step system was initially proposed in 1944 [36], and further developed in 1955 [32], by Katz and Lazarsfeld. In this system messages are not brought directly to everyone in the intended audience, but to a vocal subset who then disseminate the information among the groups they exert influence over. Figure 1.1 illustrates the differences between the one-step and two-step methodologies. Political rallies, for example, are considered to employ two-step propagation as people who attend are often opinion leaders within their communities. Research has shown evidence of the two-step system being highly applicable in real world scenarios. Norris and Curtice demonstrated that the internet was widely used for instigating two-step communications during a British election held in 2005 [47]. In their paper they argued that an increasing number of people were using two-step communications as they turned away from newspapers and news programs in favor of the internet as their primary source for information. While much of the statistics used to justify their claims are based on data exclusively from Britain, the decline of traditional news sources is evident in other nations such as the USA as reported by the Newspaper Association of America

(NAA) [50]. More recently, in 2011 Tamara Small conducted similar research into two-step communications in Canadian politics [53]. Her work detailed that people use political hashtags for spreading information collected from other areas of the internet.

Not all researchers agree that the rise of the internet as a medium for information exchange will lead to the formation of new two-step systems [9]. As news organizations, including those online, have become increasingly selective in their news reporting they have alienated users who hold opinions that differ from those being reported. This, combined with consumers possessing a disposition to seek out information they agree with [28], creates a system in which opinion leaders do not hold significantly different ideals than those in their communities. Evidence that people tend to seek relationships with like-minded individuals strengthens this hypothesis [7]. In this scenario new opinions do not often diffuse through communities as any opinion introduced is likely already held by the majority. Though it is difficult for contrary beliefs to propagate, it is common for messages that are aligned with the opinions of the majority to spread easily through these communities. These messages may even spread more efficiently than they would in a diverse community as they would be less likely to encounter resistance from users who disagre with them.

Since the main method of message propagation within the Twitter network is retweets, the network primarily leverages the two-step ideology. Retweets allow Twitter to excel at spreading information through active community members. The life cycle of a message does not end after the initial tweet but can be extended indefinitely as it is seen and shared by new users. This fashion of message propagation perfectly demonstrates information being distributed predominantly not by the initial source, but by people with whom the message resonates and therefore desire to share it. Understanding information diffusion in social networks is difficult due to the complexity of member relationships [64]. These difficulties inspired researchers to quantify characteristics of opinion leaders [52] and to determine what characteristics affect influence within social networks.

1.3.2 Twitter Influence Models

Access to the vast amounts of data created through the use of Twitter has enabled researchers to search for a more quantifiable definition of influence within the network. For example, Gilbert and Karahalios [26] mapped social activity to determine the strength of

connections between users. Research into influence on Twitter generally centers on analyzing the characteristics of users [26] or of the content of their tweets [16]. Through these types of models researchers have been able to provide rough predictions on how likely a user or tweet is to be retweeted.

User Influence

The premise of user based influence models is that there are measurable differences between users that are more likely to have their messages retweeted versus users that are more often ignored. These differences can lie in the way users are placed in the network, how they interact with the network, or who the user is outside of the network. Karthik Subbian and Prem Melville ranked users based on key metrics, including number of followers and past retweets, to determine how influential a user is [57]. Their findings indicate that the main contributing factors to influence within Twitter are number of: followers, distinct past retweets, and people mentioned by or mentioning the user. It is expected that the number of followers a user has (out-degree of a node) should influence the likelihood a user is retweeted as large numbers of followers simply creates a larger initial exposure for any tweet. However, it is noteworthy that the number of followers provides the weakest predictor of influence as compared to the rest of the variables mentioned. Subbian and Melville determined the strongest predictor of influence to be the number of times a user had been retweeted in the past, implying that the best way to be influential in the Twitter network is to have previously been influential.

Meeyoung Cha et al. took an additional step by determining influence on individual topics [16]. Looking into which users were retweeted and mentioned during a time when three different topics were popular areas of discussion on Twitter (specifically Iran, the H1N1 outbreak, and the death of Michael Jackson) they were able to compute the influence of users on each topic. They found that the top users were able to maintain their disproportionately large influence across a variety of topics. Similarly to Subbian and Melville they also determine that a user's follower count only shares a weak positive correlation with how much influence they have over the network. Cha et al. concluded that influence is almost never gained suddenly, but is slowly built through great effort and over a long period of time.

A common drawback of user based models is that they do not provide any conclusions on how influence is gained. Both papers discussed here demonstrate that the best metric

to measure a user's influence is their previous influence but are unable to provide definite reasoning as to how it was initially gained.

Content Driven

In addition to the properties of users within the Twitter network, the content of their tweets has also proven to be a reliable source of predicting how far messages will propagate. Content driven analysis of tweets has recently attained popularity; Suh, et al. found that URLs and hashtags contribute to retweetability [58]. Their work stems from the efforts of Zarrella who demonstrated that a significantly larger proportion of retweets contain URL's than non-retweeted messages [67]. Zarralla's content analysis also determined which words were the most likely to be retweeted and the reading grade level used in the retweets. He found that retweets generally have more syllables per word, more punctuation, and use more uncommon words than compared to their non-retweeted counterparts. Suh confirms and builds upon Zarrella's findings by examining metrics such as hashtags and URL domain. Suh's work indicates a strong correlation between URL domains and retweet rates. Tweets containing specific news sites such as The Onion and the New York Times received more retweets than those with links to Google News and Yahoo News. Content analysis compliments models based on the properties of users by providing suggestions on message content to non-influential users that may increase the visibility of their messages.

A combination of user and content driven models provides a solid basis upon which to predict the extent a tweet will propagate within Twitter. However, neither user nor content based models are able to provide an explanation as to how influence is gained or provide meaningful steps that users can take to become more influential. Our research shows characteristics that users can leverage and provides a model of group based influence that, when used in conjunction with the discussed methods, details a more complete understanding of message propagation within Twitter.

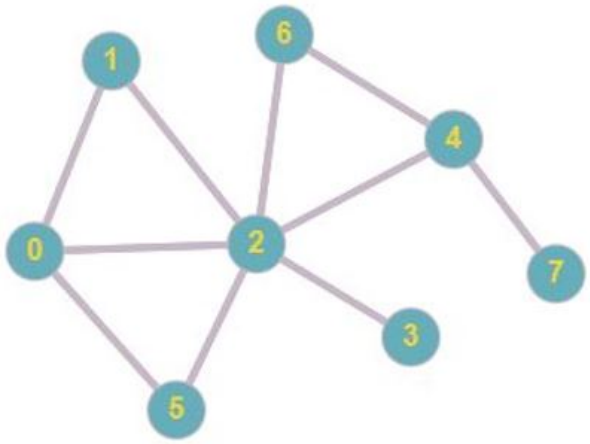
1.3.3 Network Centered Influence

Determining influence from a purely network based standpoint has long been a topic of research. There are many different metrics for measuring how central any node is in a given network, each metric providing a slightly different perspective of influence. In this thesis

we focus on Eigenvector Centrality and Expected Force as they both determine a node's influence based on that of its neighbors, making these measures well suited for detecting groups of important nodes.

Eigenvector Centrality and Expected Force are specially suited to measure influence or opinion diffusion as they assume that a node affects all of its neighbors both uniformly and simultaneously [13]. While the relationships on Twitter are not truly uniform, connection strengths (i.e. edge weightings) do not inherently exist within its network. Any weightings are due to outside influences such as real world friendships or name recognition. Eigenvector Centrality and Expected Force are therefore well equipped for developing an understanding of influence based purely on Twitter's network structure.

Eigenvector Centrality



Node	0	1	2	3	4	5	6	7
Eigencentrality	0.480	0.381	0.600	0.196	0.327	0.381	0.301	0.105
Expected Force	1.345	0.731	4.599	0.350	1.647	0.731	0.765	0.346

Table 1.1: Eigenvector Centrality and Expected Force for a sample graph (above). Node 2 is clearly central to most of the network and this is reflected by its high relative values in both metrics.

The idea that a node's influence is relative to that of its neighbors has been applied in many fields through Eigenvector Centrality [11, 39]. To calculate Eigenvector Centrality for

a general graph G(V,E) we define A be its corresponding adjacency matrix where $A_{i,j} = 1$ if v_i is neighboring v_j and $A_{i,j} = 0$ otherwise. A node's centrality relative to the rest of the network, c, is given by Equation 1.1 which can be rewritten as the eigenvector equation. Since this equation has many possible solutions the requirement that the eigenvector must be non-negative is added. The result is that only the greatest eigenvalue is able to generate centrality values. When applied to the sample graph seen above Table 1.1 we see that node 2, which is clearly the most central, is given the highest ranking. Nodes 3 and 7 have the lowest values as they have few direct connections, but 3 is significantly higher since it is connected to a more influential node (2).

$$c_i = \frac{1}{\lambda} \sum_{j \in G} A_{i,j} c_j \tag{1.1}$$

Expected Force

The Expected Force of a vertex v represents the influence that vertex exerts if it transmits a message. This is calculated by first enumerating all of the possible paths for a specified number of transmissions (n), generally $n = 2$. We could increase the value of n, but this usually yields little additional information [35]. In our sample network if $v = 7$ and $n = 2$ we would obtain the sets $S = \{[7, 4, 2], [7, 4, 6]\}$. We then define d as the number of outgoing edges between sets in S and the rest of the network, in this case $d = \{6, 2\}$. The Expected Force of v_i can be approximated by the entropy of d or

$$F(v) = -\sum_{i=1}^{|S_v|} d_i log(d_i) \tag{1.2}$$

Applying this equation to our sample network we see that Expected Force and centrality closely agree on the relative importance of the vertices. Node 2 is still dominant while nodes 3 and 7 are the weakest. For this measure however, we see that the values of 3 and 7 are nearly identical. This similarity is likely due to vertices 3 and 7 being fewer than n hops away from the central hub, 2. If we were to expand this network to include a vertex between 2 and 4 we would see a significant drop in the force of 7 as it would not be able to reach node 2 at $n = 2$.

1.3.4 Group Mentality

In order to understand group influence we must first understand why groups exist within Twitter. Whether social norms influence individuals to conform to a group mentality [23] or because people tend to be friends with others who hold similar beliefs [18, 41], groupings of like minded individuals exist both in real life and within social networks. Java et al. explored the existence of these groups, or cliques, within Twitter [29] and determined via the Clique Percolation Method (CPM, discussed below) that many overlapping communities exist where members share a common interest. An investigation of key words in tweets that propagate through these communities revealed that connected groups generally discussed similar topics. For example, Java et al. recorded a cluster of several groups that center around the discussion of technology. Each community is distinctly separate from one another, but share a common connection to the tech geek blogger, Scobleizer. While these communities have similar interests a combined analysis of the network structure and the distribution of words used demonstrated that each fills its own niche within the overarching topic of technology.

Java et al. also proposed that users join networks such as Twitter for one of three reasons: to obtain information, to spread information, or to build/maintain a friend base. More importantly they also concluded that people seek to be connected to others with similar intentions. That is, the cliques mentioned earlier are likely to be formed by groups of people who mainly fall into one of the three mentioned categories which implies that well clustered communities exist in the Twitter network that primarily wish to spread information.

1.3.5 Clique Identification

Cliques within a network are defined as a subgraph where every two distinct vertices are adjacent. They are generally categorized as maximal, indicating that they cannot be extended to include any neighbors of the current subgraph, or as k-cliques, which are complete but limited to k members. The identification of cliques has been studied in great detail [10, 60, 40]. The clique problem is NP-complete [31] so any exact solution will take exponential time with respect to the number of vertices. Since most methods have long run times any optimization is extremely useful.

Clique Percolation Method

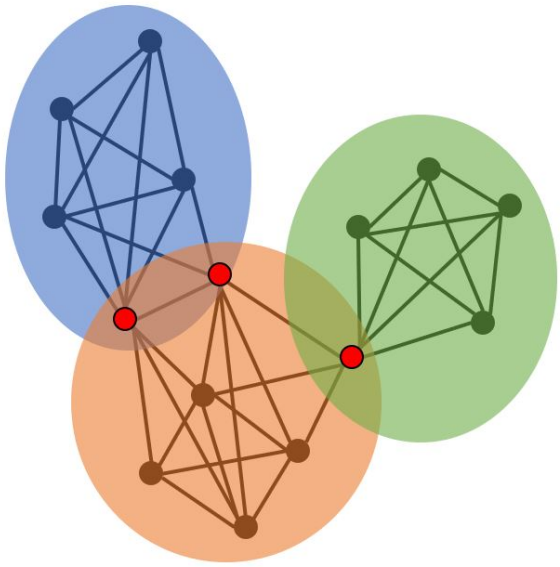


Figure 1.2: Illustration of the CPM where k=4. There are three communities, any nodes contained in multiple communities are highlighted in red.

From previous discussion we know that communities exist within the Twitter network however we have not yet discussed their underlying structure. The Clique Percolation Method (CPM) was proposed by Palla et al. to identify well clustered communities where not every vertex in the subgraph is adjacent [48]. When applying CPM every k-clique contained within the network is found and then any adjacent cliques are combined. Adjacency is defined as two k-cliques sharing k-1 nodes. Communities formed through these combinations represent the maximal union of k-cliques. Figure 1.2 illustrates the result of CPM when identifying k-cliques where $k = 4$ on a sample network. The nodes contained in multiple communities are highlighted in red. Altering the value of k has significant effects on the results of the CPM. As k decreases communities become larger to the point where at $k = 2$ every connected section of the graph will be a distinct community. Increasing k causes communities to become smaller and more disjoint until there exist no cliques large enough to satisfy the requirements of CPM.

In Twitter these communities have two important properties: 1) a natural tendency to

overlap and 2) a resistance to changes that happen outside of their respective subgraphs [29]. We also know that overlapping communities tend to share similar interests and that some of these communities exist with the desire to spread information. Combining these ideas we begin to form a picture of how these communities could be extremely important in the propagation of information and why they should be prioritized when trying to maximize message visibility.

1.3.6 Bandwagon Effect

The diffusion of innovations and ideas throughout a social or professional network is propelled by a number of factors, one of which is social pressure or the bandwagon effect [27, 21, 33]. Abrahamson and Rosenkopf explored the bandwagon effect in the diffusion of new innovations through technical communities [2]. They divided the cause of the effect into three main classifications: increasing returns, learning, and fads. For example, as groups adopted an innovation that generated profits it puts increasing pressure on other profit seeking entities to do the same. Even though Twitter is generally not used to spread technical knowledge there are certainly parallels that can be readily drawn. We discussed earlier that communities exist in Twitter with the express intention of seeking and spreading knowledge. Twitter enables these groups to increase knowledge by spreading ideas such as news stories and opinions [34]. Mendoza et al. reinforced the idea that many people use Twitter as a learning platform by stating rumors are more likely to be questioned causing them to propagate through the network differently than veritable news [42]. Since people possess a desire to learn, new knowledge in the Twitter network can be seen as akin to new innovations spreading through technical communities.

Additionally, Abrahamson et al. demonstrated that innovations are most commonly adopted through a *trickle-down process* [1]. Members of a core cluster within a social network are usually responsible for triggering adoption by outlying members, while the reverse is less likely. This process directly mirrors the two-step method as the flow of information is from community leaders to the peripheral members.

1.4 Our Contributions

Prior research reveals that determining influence is a multi-faceted problem where any one solution is likely to only capture a subset of the relevant features. Despite the success of user and content based models we believe that their results indicate they are missing an important component of identifying influence. We believe a portion of this missing piece is community based influence. Related works have shown this to be an important aspect of social networks yet to our knowledge it has not been applied to Twitter. It is clear that there are clusters of like minded users within the Twitter network. The fact that a subset of these communities actively desire to spread information indicates there is significant reason to believe that these groups play an important role in the propagation of information through the *trickle-down process* or two-step method.

Our work consists of a combined analysis of real world data and simulations designed to emulate Twitter. Twitter data is used to validate the hypothesis that influential groups exist and that they are able to propel messages through the network by increasing the rate at which messages are retweeted. Unfortunately, Twitter limits the amount and completeness of information available through its Application Programming Interface (API) making it infeasible to completely base our research on real world data. Therefore, we generated simulations to provide an environment in which we have access to a much larger set of complete information, and to provide reasonable computational times. Many of the variables that the related literature focus on, such as the strength of user connections, content of tweets, and topic specific influence were intentionally not represented in these simulations. We demonstrated that even without these variables our simulations mirror the characteristics of tweet propagation in the real Twitter network. Our accurate representation implies that our model is able to capture important driving factors in message propagation even without these variables.

With these simulations we were able to correctly identify groups of users that exert a large impact on tweet visibility and quantify their characteristics. This data was used to validate our hypothesis that targeting the identified groups significantly improves message propagation.

Chapter 2

Methods

In this section we detail the methods employed in our research. We compiled data both from the Twitter network and through simulations designed to emulate it. To obtain our real world data we traced a set of messages retweeted between 500 and 2000 times through Twitter and recorded the users who interacted with them. For our simulations we generated a network of 100,000 nodes with characteristics mirroring those found in our real world data, specifically we modeled the in/out/mutual degree and clustering of users after the Twitter network. Through analysis of real world data and simulations we developed a method to detect when large amounts of influence are exerted on the network and to extract the features of nodes in these time periods. We then detailed how we use these features to detect influential communities in new networks.

2.1 Twitter Data

2.1.1 Twitter API

All user profiles on Twitter are public by default. Users can change their visibility by editing their security settings so their tweets are not publicly accessible. All tweets from public profiles published within the past two weeks are available through Twitter's API. Any tweet older than two weeks is not accessible by the API, but may be retrieved using third party software. Tweets can be obtained through the API by one of two methods: (1)

by searching a stream of tweets being processed in real time or (2) through the data bank of messages from the two week buffer. Both methods allow for query based searches on prominent features, such as mentions and hashtags. For example, searching the real time stream for the query "#Travel" returns any tweets containing this hashtag as the message is published. The API endpoint returns a JSON object containing all of the available data including: the user who tweeted, the original author of the tweet (if it is a retweet), its contents, number of times its parent has been retweeted, etc. For reference, Figure 2.1 is a sample of data returned from the API with the query "Trump". Directly connecting to the API via its endpoints is a rather unwieldy process. There are a number of libraries created to act as an interface between software and the API. Python is supported by many of these libraries including Python-twitter, TweetPony, and Tweepy, the package used in our research [30, 43, 5].


```
{
    "created_at": "Mon Sep 17 20:48:43 +0000 2018",
    "id": 1472583691472583691,
    "in_reply_to_screen_name": null,
    "retweet_count": 722,
    "retweeted": false,
    "retweeted_status": {
        "retweet_count": 722,
        "retweet_id": 1234567891234567891,
        "retweeted_from": "TwitterUser123"
    },
    "text": "@TwitterUser123: BREAKING: according to Bloomberg the FBI ...",
    "user": {
        "followers_count": 259,
        "friends_count": 221,
        "id": 9876543219876543219,
        "screen_name": "TwitterUser321"
    }
}
```


Figure 2.1: Sample JSON data returned from the Twitter API.

To prevent Denial of Service attacks, intentional or otherwise, Twitter imposes an account based rate limit for each available endpoint. While limits are necessary for the continued existence of the API, they cause impairments in collecting large amounts of data. The rate limits are divided into 15 minute windows and once a specific endpoint has reached its limit it may no longer be accessed until the next window begins. For example, the number of requests allowed when retrieving the IDs of a user's followers is 15. Each request returns at most 5,000 user ID's, thus the maximum collection rate for this endpoint

Percentile	25%	50%	75%	95%	Max
In-degree	312	813	2290	6209	600K
Out-degree	186	574	1800	7473	3.6M

Table 2.1: Percentiles for the in and out degrees of our dataset.

is 75,000 ID's per 15 minutes. Much of the collection potential is wasted since data for multiple users cannot be obtained with the same request and most users do not have a number of followers anywhere near the 5,000 per request limit. In practice, this constrains us to collecting data on around 10-12 users per 15 minute window. At this rate it requires two days of data collection to fully trace a message retweeted 2000 times.

2.1.2 Twitter Data Set

To build our real world datasets we searched the real time stream for any messages that had been retweeted between 500 and 2000 times. Messages were located using the search query "Trump" which ensured that the tweets generally fell into the single category of politics. Since Trump is mentioned frequently on Twitter, our search was likely to yield results quickly. After a set of appropriate tweets were found we compiled the retweet history of each message as it propagated through the network within the two week buffer. While the API does not provide the ability to search directly for retweets we were able to use the body of the message as the search query which allowed for retweets to be identified with a high degree of accuracy. The results of this search represent the entire history of a single message as it traverses Twitter, including a comprehensive list of each user that retweeted the initial message. From this data the relevant adjacency matrix of the Twitter network was constructed by pulling the connections of each user in the list. The result was multiple subgraphs of Twitter each representing the life cycle of an individual message as it propagated through the network.

Our dataset was generated from 58 searches and includes approximately 45,000 users. Table 2.1 provides statistics for the in and out degrees of the users. As mentioned earlier, the degree naming convention we follow represents the direction of the flow of information. The out-degree of a node is the number of followers a user has while their in-degree is the number of users they follow.

Detecting Influence

After establishing the subgraphs from each of our samples we performed comparisons to previous research to validate the weak correlation between the out-degree of nodes and the influence they exert. Influence in other works is generally defined as the probability that a user is to be retweeted; however, due to the nature of our data every node is retweeted so this method is not applicable. In order to leverage our data we define influence as the impact that a group or individual has on the rate a message propagates. When users with large influence over the network retweet we expect to see a significant increase in the rate at which the message spreads. This measure differs from practices followed in other research in that our method is not easily able to detect the contributions of specific nodes, but instead identifies time periods in the life cycle of a tweet where a large influence was exerted.

Examining the nodes that directly precede the identified instances of large influence allows for the extraction of the features commonly found in the regions of the graph that exert the most influence. We generated our data to identify these features by randomly sampling pairs of windows from our message traces. When creating window pairs, we first generated user windows U by taking a number of users n_u who retweeted consecutively. For each u_i, a corresponding window r_i was created to contain all users that retweet within time t_r after the last user in u_i, this is illustrated in Figure 2.2. After the window selection process was complete the maximum and log-normalized average values for the out-degree were computed for the users of each user window. We compared the maximum and average values using a Spearman correlation test to the rate of message propagation in r_i.

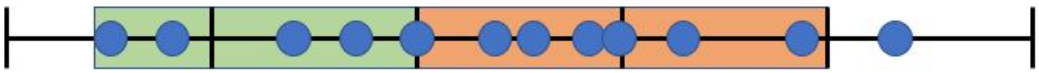


Figure 2.2: Example of a window pair for $n_u = 5$ and $t_r = 60$ minutes. Each vertical tick represents half an hour and each circle a node. u (green) begins at the first node and continues to include up to the fifth. r (orange) begins *immediately after* the fifth node and continues for one hour and in this case contains the next six nodes.

Log-normalization was applied to the averages since we desired to prove the existence of groups whose degrees were larger than average but also significantly smaller than the

largest nodes in the network. The normalization caused groups of these smaller nodes to dominate the averages as the effect of the large outliers was almost completely removed.

When looking at the average value, adjusting n_u allowed us to scale the amount to which our results were influenced by the degree of individual users as compared to that of groups. Increasing n_u decreased the resolution at which we observed the data thus diminishing the impact of individual nodes. The opposite is true for the analysis using the maximum out-degree of windows since decreasing the resolution ignored all but the largest nodes. Increasing t_r allowed us to measure the duration influence persists. Through the manipulation of these variables we were able to distinguish the impact of large singular nodes from that of groups of smaller nodes. Based on previous research, we expected to observe positive correlations between both the maximum and averaged values and the retweet rate across the board as it has been well established that more people seeing a tweet increases the chances that it propagates. Comparing the differences in the correlations of average and maximum values allowed us to quantitatively differentiate the impact of individuals versus groups.

2.2 Simulated Twitter

Gathering data purely from Twitter lead to limitations created by API restrictions. Factors including blocked accounts and information traveling through indirect connections precluded obtaining the complete history of a message as it propagated through the network. Our research only uses retweets to track message propagation, ignoring other factors such as *liking* and *commenting*. While it is possible to track comments on messages with relative efficiency there does not currently exist a method to track likes using the Twitter API making it is nearly impossible to completely trace a tweet. Even if such an endpoint did exist, there is no way to guarantee that the entire subgraph of every user who interacted with a tweet had been retrieved. To account for these holes in Twitter based data we generated a network which simulates Twitter and is based on the properties we observed in our real world data. The simulations were designed to mirror the network structure and tweet propagation observed on Twitter so that it generated data exactly as Twitter would except without the mentioned gaps.

2.2.1 Degree Distribution

To accurately simulate Twitter we needed to first understand the distribution of nodes within the network. Work done by Myers et al. suggests that the distribution of the out-degree of US users follows a power function with the 25^{th}, 50^{th} and 75^{th} percentiles being 4, 20, and 89 respectively. These values differ by several orders of magnitude from those seen in our data (see Table 2.1). This discrepancy is consistent up to the 95th percentile in both datasets until the very largest nodes, which are of similar proportions. Our papers covering different images of the Twitter network is a likely explanation for this observation. The data collected by Myers et al. represents a snapshot of the entire network at the time it was written. The data used in their paper includes active and inactive users alike. Our research focused on users that were active within two weeks of data collection. The difference between the statistics of the entire network vs the active section is staggering as the vast majority of nodes seen in our real world data fall within the top 95th percentile of all nodes in the Twitter network. This variance implies that much of Twitter consists of small dead nodes.

If these dead nodes are uniformly intermixed with active nodes then the distribution of the degree of nodes in our simulations should match the description in Myers et al. If, however, these nodes are not generally connected to the active section of the network then we must design our simulations to more closely align to what we observed in our Twitter dataset. To make this determination we sampled followers of the users in our real world data. The resulting distribution of their degrees very well aligned with the data we observed from Twitter indicating that the degree of nodes in our simulations should as well.

2.2.2 Network Generation

Without access to a powerful cluster of machines the full Twitter network is too large to be simulated efficiently. To keep run-times within reasonable time frames our simulations were limited to 100,000 nodes. We generated our networks by first creating every node and internally specifying an exact number of followers and an approximate number of folowees. We calculate a node's followers and folowees by sampling a random value from $[0, 100]$ to use in Equation 2.1. Values for followers and folowees are both generated from this equation, but each uses a different set of constants; $a = 15.267, b = .065, c = 150$ and

$a = 24.548, b = .059, c = 250$ for followers and followees respectively. At the lowest values the equations yield a smaller number of followers than folowees until a threshold is reached and this is reversed. This is in keeping with both our findings on the relative degrees of nodes and the findings of Myers et. al.

$$F = ae^{-bx} + c \tag{2.1}$$

Once the distribution of followers and followees has been calculated for every node, they are each assigned an approximate clustering value. The assigned value is the desired clustering value for a node once all the edges have been generated. Drawing again from the work done by Myers et. al we know that a node's clustering value is dependent on its degree as nodes of smaller degree tend to be more tightly clustered with their neighbors. With this process we are able to create a set of nodes each knowing their expected in/out-degree and clustering values such that they closely mirror Twitter. Edges for the network are generated by Algorithm 1 (located at the end of this chapter).

The edge generation algorithm incrementally assigns followers to nodes until there are no longer any nodes that need followers. At the start of every iteration our model selects follower candidates from one of three groups: all nodes exactly two steps from the current node, all nodes that the current node follows, or the entire network. The first option causes the generation of clusters within the network and is selected with a probability equal to the expected clustering of the node currently being evaluated. Nodes with higher expected clustering will more frequently select followers from this set. The expected clustering of a node n is given by Equation 2.2. The second option represents the likelihood that a node will follow people who are already following them. Selecting nodes from this group will increase the *mutual degree* of the network. The smallest nodes in our network are set to each have a mutual degree of around 50% of their out-degree. This value scales down to 25% as the nodes increase in size.

$$C_n = Max(.3 - \frac{log_{10}(n_{out})}{20}, .1) \tag{2.2}$$

From the selected candidates we preform a weighted choice to determine who should follow our current node. The weight generated for each node is equal to the proportion of its current in-degree compared to the value we expect it to attain after all the edges are generated. As nodes follow others they become less likely to be selected again.

Percentile	25%	50%	75%	95%
In-degree	372	713	2138	6848
Out-degree	227	543	2150	7491

Table 2.2: Statistics for the in and out degrees of our simulated dataset.

Many efficiencies were developed for this algorithm in an effort to decrease computation time while having no impact on results. The most time expensive operation throughout the process was to compute the neighbors and intersections of multiple nodes; these steps are taken only when necessary. For example, we postpone validating that a candidate is not already following the node we are evaluating until the selection has already been made. It is more efficient in the long run to allow all possible connections and retry if an invalid choice is made than to filter choices and remove the chance for duplicate connections. We also avoid checking the effect that connections will have on a node's clustering unless absolutely necessary. Table 2.2 shows the percentiles of in/out-degree for our generated network which we observe are comparable to the values in our Twitter data.

With all of our edges and nodes generated we created a network of 100,000 nodes that closely resembles Twitter in both degree and clustering distribution. By sending tweets through our simulated networks we generated our own complete sample data in significantly larger quantities than was available through Twitter's API.

2.2.3 Tweeting in the simulation

After either a randomly or intentionally selected node injects the first tweet into the simulation it propagates in a series of iterations. Within each iteration there exists independent lists of the nodes that have tweeted, nodes that have seen the tweet but not yet tweeted, and those that have not had contact with a the tweet. We generated a probability of tweeting for every node in the second list given by Equation 2.3. Nodes in the first and third lists cannot tweet.

$$P_n = P_{base} * d_n * sp_n \tag{2.3}$$

The equation contains three main factors that influence a user's likelihood to retweet a message. The first being a base probability shared by every node which represents the

probability that any user will retweet without outside influence, P_{base}. We also introduce a decay factor d for each user so that the probability to retweet decays the longer it has been since the last time the user saw the message. This decay resets every time a node that a user is following retweets the message. The last factor sp, drawn from the bandwagon effect discussed earlier, introduces social pressure. Users that see a message from multiple sources are more likely to retweet it.

During every iteration each user decides if they are going to retweet based on their own probabilities. Once each decision is made, those following users who retweeted are considered to have seen the tweet and the lists are updated to reflect the new state of the network. The cycle repeats until a termination condition is reached. We end these runs when no additional users have retweeted in a specified number of iterations (we set this to be 5). Once a run through the simulation network is complete the path the tweet takes is recorded and properties of the nodes involved are analyzed.

Tuning the Simulation

The balance between the base probability that any user retweets and the social pressure they exert on the network is highly sensitive. Simulations show that even slight changes to either of these values can dramatically impact the way the network operates. Runs where the base probabilities were set too high or too low resulted in either the tweet rapidly spreading over the entire network, or never being retweeted at all. Decreasing the social pressure variable caused tweets to grow logarithmically until a portion of the network relative to the base probability had tweeted. When the social pressure was set too high we saw explosive growth generated from just a few users tweeting. Once this began the entire network was quickly consumed as every node rushed to jump on the bandwagon. Through careful adjustments of both parameters we created a system where users were able to exert significant influence on those around them but were not able to drive the entire network. We arrived at the values given below where r is the number of iterations since the decay has been reset for a node and s is the number of sources a node has seen the tweet from.

$$P_{base} = .002, \quad d = \frac{1}{3r}, \quad sp = \frac{s}{500} + 1 \tag{2.4}$$

2.2.4 Identifying Clusters

Influence in our simulated runs was measured with the same method employed on our Twitter data. Since simulated data provided access to the complete network structure we were able to identify cliques and directly measure their contributions. To do so we first located groups of nodes that appeared close to the large spikes in the retweet rate. Employing Eigenvector Centrality, Expected Force and CPM we determined how central these nodes were in our network and the structure of the subgraphs surrounding them.

CPM is exponentially expensive to compute with respect to the number of vertices so we only performed the algorithm on the set of nodes we identified within influential areas. We applied CPM on those areas for progressively smaller k values to better define community structure. Since communities are independent of the network outside of their own respective subgraphs this allowed us to understand the characteristics of the groups of nodes directly linked to increasing message propagation without having to analyze the entire network. We filtered out any groups of nodes that were not well clustered and then compared the centrality and force of nodes in the remaining communities to a random sampling from our network. The difference between these sets defines the characteristics of influential and non influential groups.

Using message propagation data to identify influential communities prevented us from detecting these clusters before messages spread through the network. However, we designed an algorithm to preemptively identify important communities by highlighting areas in a network with properties similar to influential communities in our simulations. These communities are characterized by a high Expected Force, Eigenvector Centrality and being well clustered. Through this process we were able to provide insight on which network areas are the most influential when the only available data is the network's structure.

```
Data: Nodes each given their expected in/out-degrees and clustering values
Result: Complete network with all edges generated
1  while There are still nodes who still need followers do
2      for n in Nodes do
3          if n does not need a follower then
4              continue
5          end
6          r = newRandomFloat;
7          choices = Nodes − n;
8          if r ≤ n.expectedClustering then
9              choices = n.getTwoStepConnections
10         end
11         if r > mutualProbability then
12             choices = n.getIsFollowing()
13         end
14         choice = weightedChoice(choices);
15         if choiceIsValid(choice) then
16             choice.follow(n)
17         else
18             goto : 15
19         end
20     end
21 end
```

Algorithm 1: The algorithm iteratively assigns followers to nodes based on a weighted choice which is more likely to select candidates that are furthest from the number of nodes they are expected to follow. At the start of every iteration it is randomly decided that candidates will be pulled from one of three groups. These being: the whole network, all nodes exactly two steps from the current node, or all nodes that the current node follows (the current node being n in algorithm). The larger the *expectedClustering* or *mutualProbability* the more likely their respective groups are to be selected.

Chapter 3

Results

Through the application of methods described in the previous section we determined how to detect influential communities and their relative importance. When we pulled our data from Twitter we observed regular times where the rate that a message was retweeted would drastically increase (see Figure 3.1). Determining the causes of these points was our initial focus. In this section we first establish that these inflection points can be caused by tight knit communities. We then generate simulation networks and verify that they accurately capture the same activity as we observed in our Twitter data. Through these simulations we identify characteristics common to influential communities in the generated networks. Finally, We create an algorithm that leverages the identified characteristics to locate influential communities in new networks. To justify our algorithm we demonstrate that prioritizing spreading messages to areas we identify generates higher message visibility when compared to individual nodes with high centrality/force.

3.1 Detecting Influence

We first wished to establish the relative influence of groups and individuals within Twitter. Using the method discussed in section 2.1.2 we determined which properties were common to nodes surrounding spikes in retweet rate. Previous work indicates that there should be a weak correlation between these spikes and the out-degree of nodes around them. In the following sections we preform analysis on both our real world data and our simulations.

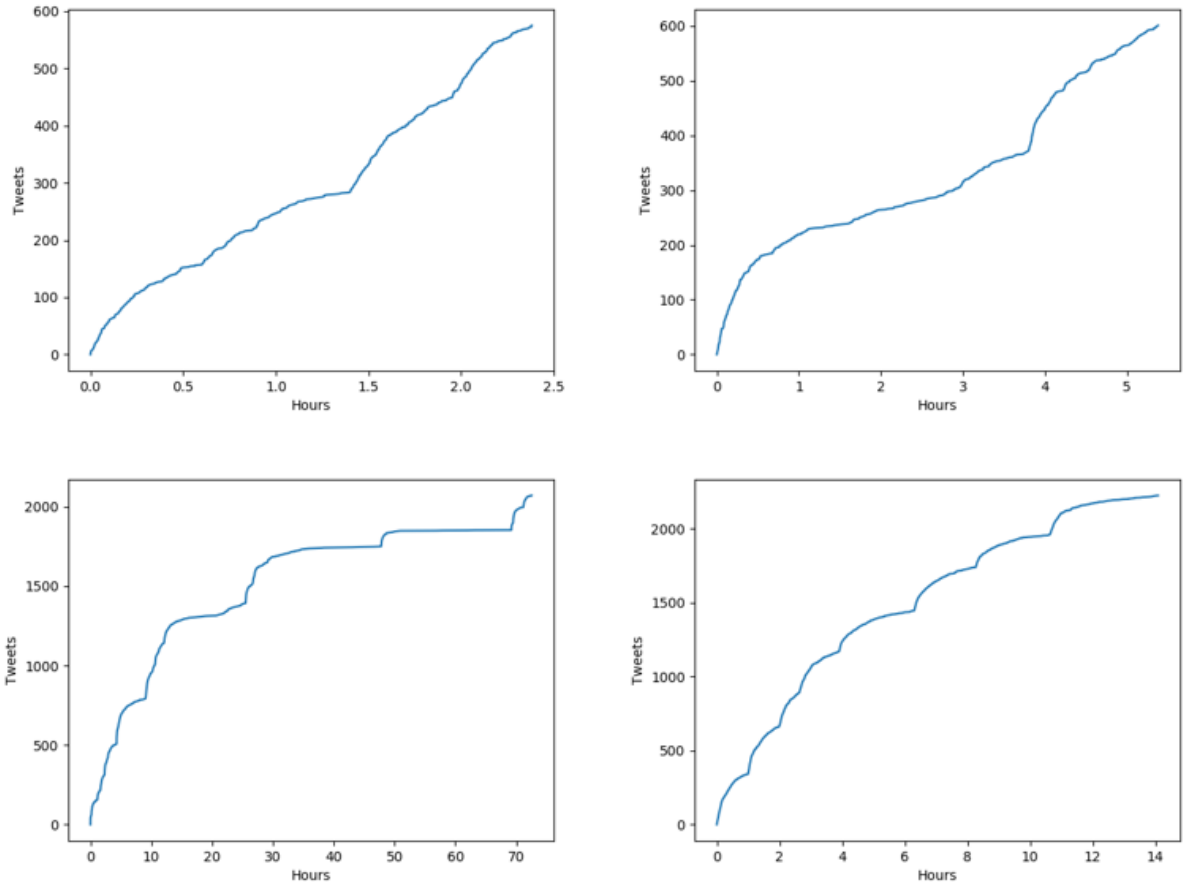


Figure 3.1: Sample traces from our Twitter data.

Since we designed our simulation to mirror Twitter we expect to see similar results when the same analysis is performed on both datasets.

3.1.1 Real World Data

A sample from our Twitter data, depicted in Figure 3.2, illustrates the number of followers for users contained in a tweet's life cycle (left) and the number of times the message was retweeted (right). There are clear inflection points where propagation significantly increased indicating that the tweet encountered an area in the network with a high degree of influence. At $t \approx .9$ we see a node that is an order of magnitude larger than any other in

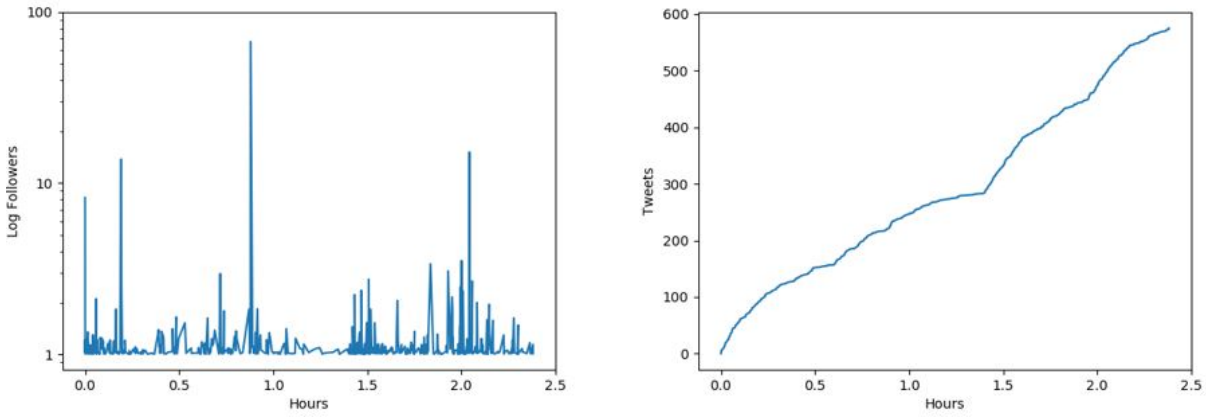


Figure 3.2: A sample tweet from Twitter. The number of users who have tweeted (right) and the corresponding number of followers for the tweeting nodes (left).

this set, however it appears to have very little effect on the message propagation. Instead, we find that the two spikes in retweet rate are located near a group of significantly smaller nodes at $t = 1.5$ and a different large node at $t = 2$. While the nodes at $t = 1.5$ are larger than average we often refer to groups like these as "small nodes" as we are generally comparing their influence against that of the largest nodes in the network.

Influence of Degree

We sampled sets of consecutively tweeting nodes from our data and preformed a Spearman correlation test to compare the out-degree of the user with the most followers in each set to the rate at which tweets propagate immediately after. The results, illustrated in table 3.1, demonstrate the expected weak correlation ($p < 0.05$ for all values) between the maximum size of nodes and retweet rate. As n_u was increased, meaning each set included more users, we observed a higher correlation between the two variables indicating that processing small nodes generates noise and that extracting only the largest nodes yields a cleaner signal. This supports the conclusion arrived at by other researchers that the out-degree of individual nodes plays a small role in determining influence.

To differentiate between the influence of the largest nodes and clusters of smaller ones we log-normalized our follower data. The groups of nodes we analyzed with the Spearman correlation test are larger than average thus our analysis with the maximum value included

Degree	$n_u = 5$	$n_u = 10$	$n_u = 25$	$n_u = 50$
$t_r = 5$	.0321	.0462	.0477	.0532
$t_r = 30$	.0307	.0382	.0346	.0490
$t_r = 60$	.0263	.0336	.0323	.0482
$t_r = 120$	.0342	.0403	.0360	.0571
$t_r = 5$	.0176	.0271	.0342	.0331
$t_r = 30$	.0169	.0228	.0292	.0329
$t_r = 60$	.0126	.0185	.0248	.0284
$t_r = 120$	.0209	.0266	.0320	.0387

Table 3.1: Spearman correlation coefficient comparing different n_u and t_r values using the maximum (top) and log-normal average (bottom) follower value for each window. n_u and t_r control the number of consecutively tweeting nodes we observe per window and the duration (min) that we record the retweet rate respectively.

any correlation due to these groups and from very large nodes. Log-normalization rescales the largest nodes to be much closer to the average, thereby negating their impact on the correlations. We have, in essence, subtracted the effect of single large nodes from our previous measurement. Any remaining correlation is largely due to groups of smaller nodes.

Although the correlations based on log-normalization are slightly weaker for all values of n_u and t_r than when using the maximum, the fact that a positive correlation still exists implies that groups of smaller users tweeting together does impact message propagation. If this were only a property of extremely large nodes we would expect correlation values to drop near zero. These results indicate that while single large nodes can have an impact, large increases in retweet rate are also found near groupings of nodes with smaller degrees.

Group Closeness

We have validated that groups of smaller nodes tweeting together can lead to an increased retweet rate, but we have not yet demonstrated any relationship between these users. We once again sample window pairs from our runs to capture the effect of cliques spreading the tweet among themselves, but we alter the time window to begin with the first tweet in its corresponding user window. Comparing the average shortest path between every

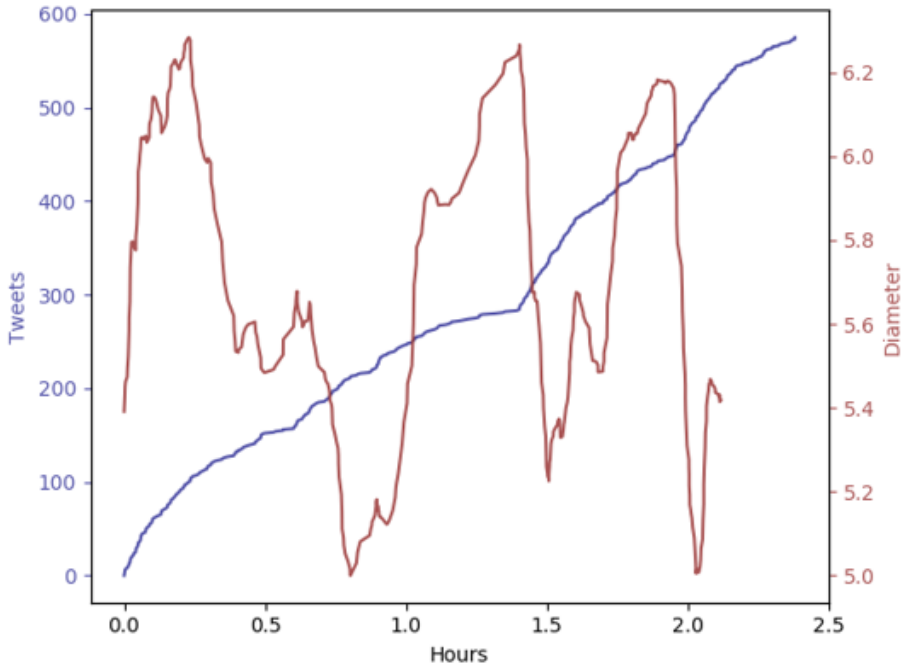


Figure 3.3: Graph of the average shortest path length for each window overlay with the number of retweets.

pair of nodes in the subgraphs found in each user window to the retweet rate in their corresponding time window revealed how tightly knit the nodes contained in these groups were. Due to Twitter data limitations there are many nodes that are disjoint from the rest of our network since we were unable to collect the complete connection data. If we encounter disjoint nodes in our windows we defaulted to a value equal to the maximum diameter found across all subgraphs. As the maximum diameter was not significantly larger than the overall average, this assignment allowed for subgraphs with few missing paths to still have a low average path length while ensuring subgraphs with mostly missing paths did not. Figure 3.3 illustrates the averaged diameter overlay with the number of retweets for one of our traces. We clearly see that the average path length falls sharply at both of the inflection points indicating at those times the tweet was spread through groups of closely connected nodes. Table 3.2 shows the overall correlation between average diameter and the retweet rate.

Shortest Path	$n_u = 5$	$n_u = 10$	$n_u = 25$	$n_u = 50$
$t_r = 5$	-0.146	-0.162	-0.162	-0.169
$t_r = 30$	-0.228	-0.257	-0.255	-0.268
$t_r = 60$	-0.251	-0.283	-0.282	-0.296
$t_r = 120$	-0.280	-0.315	-0.318	-0.333

Table 3.2: Spearman correlation between average diameter and retweet rate.

These correlations are significantly stronger than those we observed when comparing the out-degree of nodes. The negative correlation demonstrated here, along with the our results from analyzing the out-degree of nodes, provides sufficient evidence that influential groups not only exist within the Twitter network, but that members of these groups are closely connected to one another. Our next step would be to observe the network structure within these groups, but this analysis proves to be impossible due to the missing sections in available Twitter data. We leverage simulations to account for these gaps.

3.1.2 Simulation Data

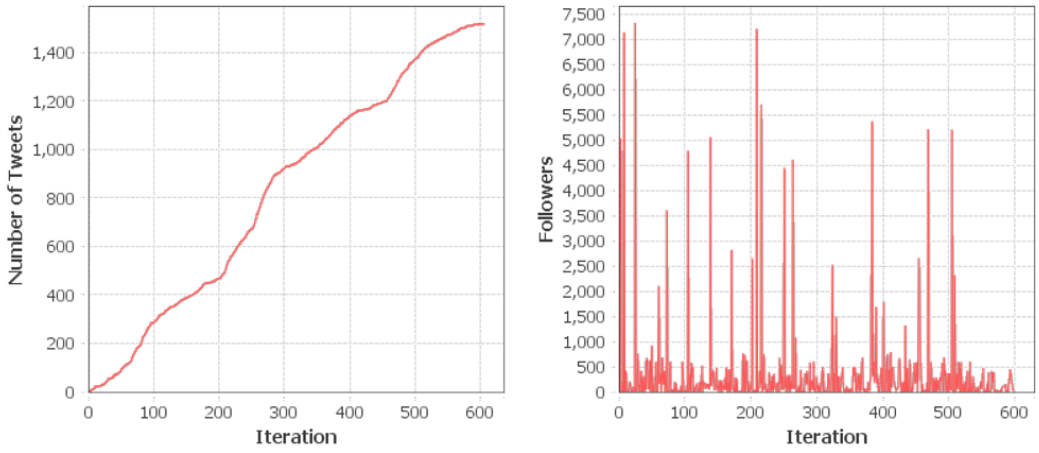


Figure 3.4: Sample run from our simulations detailing retweet rate (left) and follower count (right).

We repeat the analysis done using real world data in the previous section with our

simulated network to confirm we have accurately recreated the characteristics observed within Twitter. Figure 3.4 represents a sample run from our simulations with the number of retweets (left) and followers (right). Just as with the data from Twitter there are clear inflection points where the retweet rate spikes suddenly.

Any retweets that occur within the same iteration do not have an order as they occur simultaneously within the system. We are able to assigning a random ordering without distorting results since each iteration contains few nodes and there a large number of iterations for each run. It is also not possible to assign an exact amount of real world time that each iteration represents so the values of t_r instead denote the number of iterations each window r contains. We can estimate an iteration/time conversion by comparing the number of users contained in the average iteration to the length of time it usually takes an equivalent number to tweet in Twitter. On average we observe 100 users tweeting per hour and that an iteration contains 3 to 4 users. From this we approximate 25 iterations to be one hour of real time.

Table 3.3 outlines the correlations between both the maximum and log-normalized averages of degree and retweet rate for our simulations. As with our Twitter data, these results demonstrate that out-degree plays a small role in determining influence in our simulations for both groups and individuals. When compared with the results from the previous section (see table 3.1) we observe that although increasing both t_r and n_u has more impact in our simulations than with the Twitter data the two measurements generally agree. This indicates that while out-degree is slightly more of a driving factor in our simulations it is not significantly more powerful than in Twitter. Table 3.4 details the relationship between diameter and retweet rate in our simulations which also closely agrees with our findings using Twitter data.

Having validated that our simulation data provides a reasonable match for our observations on Twitter, we can justifiably use simulated runs to extract and analyze the properties of influential groups in our network.

Importance of Clusters

To prove that clustering has a major impact on message propagation we tested the rate at which messages spread in two networks: one with Twitter-like clustering and one with clustering significantly lower. To generate a network with much lower clustering we removed

Degree	$n_u = 5$	$n_u = 10$	$n_u = 25$	$n_u = 50$
$t_r = 5$	.0328	.0377	.0624	.0918
$t_r = 30$	.0363	.0434	.0709	.1019
$t_r = 60$	.0373	.0460	.0777	.1129
$t_r = 120$	.0380	.0436	.0749	.1162
$t_r = 5$	.0105	.0289	.0469	.0716
$t_r = 30$	.0134	.0252	.0535	.0750
$t_r = 60$	.0163	.0247	.0592	.0741
$t_r = 120$	.0193	.0255	.0591	.0732

Table 3.3: Spearman correlation between our simulation out-degree, for both maximum (top) and log normal average (bottom), and retweet rate.

Shortest Path	$n_u = 5$	$n_u = 10$	$n_u = 25$	$n_u = 50$
$t_r = 5$	-.1057	-.2891	-.4694	-.4169
$t_r = 30$	-.1347	-.2528	-.3350	-.3504
$t_r = 60$	-.1634	-.2470	-.3926	-.3415
$t_r = 120$	-.1930	-.2557	-.3912	-.3321

Table 3.4: Spearman correlation between our simulation retweet rate and average diameter.

lines 8-10 and 11-13 in our edge generation algorithm (1). Any network generated from this modified code has randomly assigned edges but still retains a Twitter-like distribution of degrees. Message propagation is expected to be proportionately lessened since the clustering has been significantly lowered across the entire network. To validate, we recorded a set of 1000 runs through both our original simulation and the random network and then compared the number of times the messages were retweeted. Due to the low base probability of tweeting it was extremely likely that the initial message would never be retweeted in any given run. To avoid uninformative results we required there be at least one retweet in each run.

As expected, our results in table 3.5 show significantly fewer retweets in the random network. Without the ability for groups of nodes to influence those around them a critical component of the driving force behind message spread was lost.

Additionally, we compared the clustering of nodes in a tweet's life cycle to the rate of message propagation. Figure 3.5 depicts the maximum clustering value in every iteration

and the retweet rate of the message. Even when only the max clustering from each iteration was observed we achieved a noisy, but accurate, representation of the retweet rate. Across 1000 runs the Spearman correlation coefficient averages to 0.796 with $p << .05$.

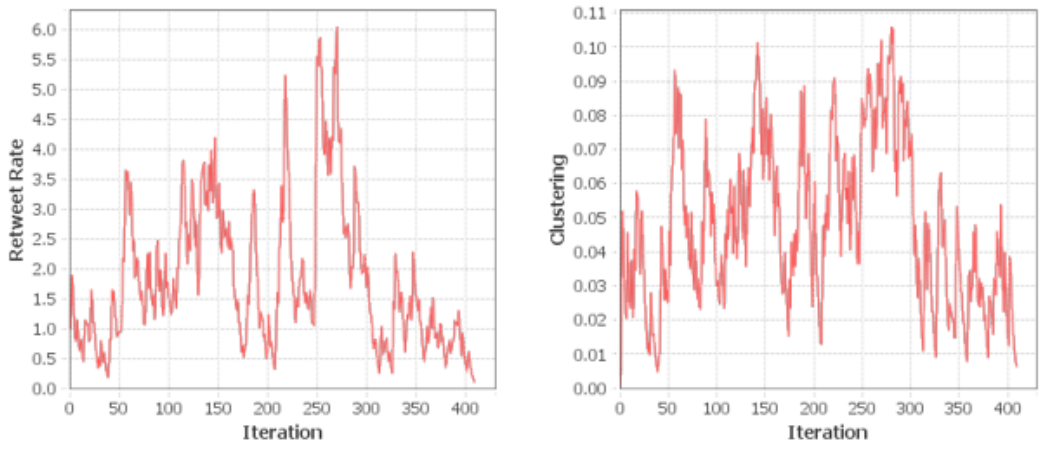


Figure 3.5: The average clustering value found in each iteration (right) compared to the (smoothed) rate the message propagates (left). Despite noise the peaks in retweet rate are clearly defined in the clustering.

We established that well clustered groups of nodes exist both in our simulation and real world data and that these clusters contribute significantly to message visibility.

3.2 Cluster Properties

Finding influential clusters when we have access to the retweet rate is trivial. Simply isolating nodes that appeared around the same time as spikes in the retweet rate gave us

Retweets	25%	50%	75%	85%	95%
Clustered	2	3	5	12	90
Random	2	2	2	3	4

Table 3.5: Comparison between the number of retweets in a random network and a network where Twitter-like clustering is enforced. The probability of retweeting is calculated identically for each network. Each run was required to be retweeted at least once.

a set containing all of the relevant clusters. We identified areas of the network containing influential clusters by when the number of nodes tweeting in an iteration crossed a set threshold. We defined three thresholds for the number of tweets per iteration: one high, medium and low, set to 9, 7, and 5 respectively. The low threshold was set to catch even a mild spike in tweets while the high threshold only triggers from the largest. To determine what differentiates these nodes in our identified areas from the rest of the network we compare their Eigenvector Centrality, Expected Force, and network structure to random samples from our simulations. The dataset for this analysis was 100 simulation runs, where each was required to have at least 700 retweets.

3.2.1 Clique Percolation Method

Eigenvector Centrality and Expected Force are good metrics to apply to entire networks; however, they do not guarantee that groups they identify are well clustered. Since these metrics operate by measuring the out-degree of nearby nodes they will indicate areas around single large nodes as influential. While this may be true, these areas are not the focus of this paper as we wish to analyze only well clustered groups of nodes. To accomplish this we generated subgraphs by sampling sets of nodes found around the spikes in retweet rates. All nodes with fewer than 3 neighbors in their set were removed and the neighbors of all remaining nodes were added. CPM was performed on each of these subgraphs at varying resolution, or k. For every spike in retweets identified we sampled a random grouping of nodes in our dataset for comparison. From this we determined the most common structure found in both random samples and our clusters.

CPM	$k = 3$	$k = 4$	$k = 5$
Communities - Window	901	191	10
Total Nodes - Window	3.39e5	8915	163
communities - Random	190	16	0
Total Nodes - Random	12945	123	0

Table 3.6: CPM for the nodes found in areas near spikes in retweets. For values larger than $k = 5$ there were no communities found. The total nodes represents the number of total nodes (including repeats) found across all 100 runs contined in communities of size k.

Table 3.6 illustrates that across our dataset there exist multiple large communities for $3 \leq k \leq 5$. We find these communities are not nearly as common in random samples as they are within our selected windows. While a large $k = 3$ community would likely be a sufficient definition for a well clustered group, we employ a stricter measure for greater accuracy. We take any region around a spike that contains a $k \geq 4$ community larger than 22 members to be an instance where the spike in retweets was caused by an influential community. After filtering communities seen more than once across our set of runs this definition yields around 70 distinct clusters, as reflected in table 3.7, containing a total of approximately 3000 nodes, or $\approx 3\%$ of the total network. We use these clusters in the following calculations for Centrality and Expected Force.

Nodes per Community	25%	50%	75%	95%
$k = 3$	10	60	517	2177
$k = 4$	9	22	50	173
$k = 5$	5	8	13	21

Table 3.7: Breakdown of the nodes contained within each k community

3.2.2 Eigenvector Centrality and Expected Force

Calculating Eigenvalue Centrality and Expected Force provides a representation of how central nodes are purely based on their network location. Using the clusters identified during the previous section we compared averages of both these values to a random sample. Our results, shown in Table 3.8, determined that the nodes contained within the windows have a slightly higher centrality, but have a significantly larger force.

The range of the values returned by both of these calculations is highly dependent on properties of the network being analyzed. Variables such as network size and degree distribution cause large fluctuations in results. For example, the highest force in the sample network at the beginning of this paper was 4.5, but here we see values many orders of magnitude larger. This causes centrality values for any node to indicate influence relative to other nodes in the network. Since the centrality of our important nodes is only half a standard deviation above average, we can discard it as a measure of influence in favor of using their Expected Force which is almost two full standard deviations larger than average.

Node Location	Centrality - avg	Force - avg	Centrality - σ	Force - σ
In Clusters	2.95e−3	$7.16e9$	-	-
Random Sample	2.01e−3	$2.12e9$	2.1e−3	$2.35e9$

Table 3.8: Average for the Eigenvector Centrality and Expected Force measurements.

With these results we have identified that influential communities are more well clustered and have higher Expected Force than the rest of the network. While this distinction was generated with the use of time dependent data, we were able to work backwards to identify influential communities in a network without this data.

3.3 Community Identification

Our research culminated in the identification of influential clusters without the use of time based information. That is, we created an algorithm to identify influential communities within a network based only on its structure. When applied to a fresh network our algorithm operates in two steps. We first generate a set of all nodes with Expected Force greater than 2 standard deviations above average. Any nodes with fewer than 2 neighbors in this set are removed to filter out isolated outliers. The set is then extended to include the neighbors of all remaining nodes yielding multiple connected components representing our candidates for influential communities. We apply CPM to these candidates to determine which ones contain the required structure which, in the previous section, we determined to be at least 22 nodes in a k-clique community where $k \geq 4$. Removing any communities that do not follow this requirement creates the remaining set of well connected communities. We rank these communities by the summation of the Expected Force of their nodes.

To test that spreading messages to communities with the highest rankings maximized message visibility we created a fresh network with an additional node, N. As we wish to show that our algorithm preforms for even average sized users the out-degree of N was chosen to be 450. Two sets of 1000 simulations were run with N as the initial tweeter. For the first set of runs we connected N to nodes contained within communities we identified as influential. For the second set N connected to 450 nodes whose Expected Force were closest to the average displayed by nodes in our clusters, not including those that were identified by our algorithm. In a third set of 1000 simulations N was removed and instead

Retweets	25%	50%	75%	85%	95%
$Alg_{r=1}$	3	5	24	73	464
$EF_{r=1}$	3	5	12	22	121
$Rnd_{r=1}$	2	3	5	8	42
$Alg_{r=10}$	19	46	237	551	1231
$EF_{r=10}$	16	33	156	416	985
$Rnd_{r=10}$	16	25	149	483	1150
$Alg_{r=100}$	207	405	886	1220	1801
$EF_{r=100}$	209	433	868	1183	1747
$Rnd_{r=100}$	199	406	898	1218	1969

Table 3.9: Percentiles for the number of times a message was retweeted. Alg, EF, and Rnd denote our algorithm, nodes with Expected Force two standard deviations above average and random start nodes respectively. r is the minimum number of retweets required for a run to be counted.

random starting nodes with an out-degree of 450 ± 50 were chosen. For all three sets we specified different minimum levels of retweet, r, to determine where our algorithm is most effective. The results are displayed in Table 3.9.

Clearly our algorithm has the greatest impact when the minimum number of retweets is kept small. For $r = 1$ we observed that our algorithm outpaces both the random set by a factor of 10 and the average Expected Force set by a factor of 4. The comparisons between the results of our sets implies that combining our identified network structure and Expected Force generates better recommendations than either do separately.

As the number of retweets increased beyond 100 the impact of our algorithm begins to diminish. We might expect that if our model makes messages propagate more efficiently to start then it would have a corresponding effect later in a tweet's life cycle. We do not see this to be the case which implies that at large numbers of retweets our algorithm does not have significant impact on message growth. That is, our algorithm is very effective in stimulating early message propagation, but after a point the network structure surrounding the initially tweeting node matters little as the effects from the rest of the network dominate the retweet rate. Our results demonstrate a working algorithm, based only on network structure, that is effective in increasing tweet visibility.

Chapter 4

Discussion

In recent years Twitter has become one of the largest online social networking platforms in the world making it a valuable resource for anyone seeking to spread a message. Billions of dollars are spent annually on social media marketing indicating the prevalence of these platforms. In this paper we have shown that by exclusively analyzing Twitter's structure it is possible to increase message visibility and therefore presence on the network.

We began by discussing that Twitter follows the two-step methodology for information diffusion. Through related works we know that communities dedicated to specific topics exist and that these communities are interconnected. We also know that some of these communities exist solely for the purpose of receiving and spreading information. Messages propagate through these communities via retweets by bouncing from user to user. In this way information flows from one active community to another instead of directly from a central source.

The Bandwagon Effect informed us that communities could generate rapid message propagation. This effect causes individual group members to influence each other in the form of social pressure and implies that nodes which are more centrally located within groups will have significantly more influence than those on the fringe.

Through analysis of the strengths and weaknesses of other models of influence on Twitter we identified a gap in the current literature concerning an applied network analysis. User based models weakly attribute influence to out-degree, but mainly to resolve that influence begets influence. Models that take into account the content of tweets arrive at

a similar conclusion while also being able to point out several key features of tweets that influence propagation, such as URL usage, hashtags, and mentions. Both methods show promise; however, neither is able to provide a complete definition of influence.

To fill this gap we analyzed the network structure of Twitter. By collecting data through the Twitter API we first reaffirmed the notion that out-degree only weakly corresponds to influence for both groups and individuals. We then displayed a much stronger correlation between the average diameter of recently tweeting nodes and tweet propagation. The results implied that clustered groups of nodes tweeting simultaneously were located near increases in retweet rate. Due to API restrictions this was the extent to which our analysis on real Twitter data was able to proceed.

We extended our data by generating simulation networks that modeled Twitter in both degree and clustering distribution. These simulations were proven to accurately mirror Twitter as we again demonstrated a weak correlation of retweet rate to out-degree, and stronger correlation to the distance between recently tweeting users. We learned from these results that many spikes in retweet rate are likely caused by groups of well clustered users all tweeting together. Building upon this we leveraged our simulation data to validate that clustering, specifically, was a main driving force behind message diffusion as lowing clustering of the network severely crippled message propagation.

Using Eigenvector Centrality, Expected Force, and CPM we identified the properties of influential clusters. We concluded that nodes in these clusters had significantly larger Expected Force and were contained within large clustered communities. Working in reverse we applied those properties to identify clusters in new networks. Finally, we developed a general purpose algorithm that can be applied to entire networks or subgraphs. This algorithm generates priorities for connecting to sections of the graph when striving to increase message visibility. After testing the algorithm on a newly generated network we found that our algorithm primarily increased the rate at which tweets propagate early in the tweet's life cycle while its effect diminished at higher numbers of retweets. Even with the diminishing effect, we confidently conclude that our research allows us to more efficiently spread messages through our Twitter-like networks.

The generation of this class of algorithm has immediate implications for message propagation in social networks. Many entities benefit from the ability to jump-start a tweet and quickly spread it through a network making this work highly valuable in political and advertising spheres. Moreover, this research provides a method for groups to push their

agendas by leveraging the influence of echo chambers. Recent events in the United States such as Russia spreading misinformation to influence the 2016 election and the outbreak of measles due to the anti-vax movement clearly demonstrate the potency, and danger, of manipulating social networks in this manner.

4.1 Weaknesses

Research based on the Twitter network, such as this thesis, suffers from a difficulty in obtaining large amounts of relevant data. Correspondingly, the largest threat to the validity of this paper is the lack of a concrete proof of its real world application. While we show significant supporting evidence that our model will hold up in practice, with our current resources we are unable to fully test its abilities.

This lack of real world data also makes it difficult to completely justify the accuracy of our simulations in recreating the Twitter network. While the fact that our analysis on both real world and simulation data produces similar results is generally convincing, this is not completely satisfying. A more in depth look into our choices for the values of social pressure and base probability is called for. The spread of opinions on generic social networks is well documented, but insufficient literature exists on the these factors within Twitter to make entirely informed choices.

It is not trivial to say that even with the available data the exact algorithm we developed would be directly applicable to Twitter. There almost certainly exist slight differences between our simulations and the Twitter network that would cause our methods to require fine tuning. However, it is highly improbable that our premise is far off base. Slight adjustments to the parameters we search for when identifying influential communities, such as the optimal k-clique substructure and Expected Force, would likely be necessary.

4.2 Future Work

Our algorithm clearly shows promise in theory so the most immediate addition that would benefit this research is to directly apply our algorithm to Twitter. As our algorithm is designed to locate areas of networks where tweets are likely to spread rapidly it is most

useful in identifying these locations before tweets spread to them. There is no guarantee that any small areas analyzed will be active thus it would be vital to cast a large net and cover as much of the network as possible. This would require another snapshot of Twitter as was done in the paper by Myers et al. This then causes the issues discussed earlier where work is limited by Twitter's API restrictions. Any attempt at this would require the use of multiple twitter accounts to simultaneously pull information through the API.

Integrating our method with related research after applying it to Twitter would be another possible course of action. For example, a combination of our predictions based on network structure with the content model of Suh et al. and the understanding of individual users given by Meeyoung Cha et al. could provide a powerful model. Each of these components contains a subset of what defines influence in Twitter and the combination thereof would assist in identifying any additional missing pieces.

One such missing piece of current research is the effect of bridge nodes in connecting well clustered communities. It was discussed earlier that many topic specific groups in Twitter are connected by individuals whose interests overlap with each group. Identifying the role that these users play in spreading messages could further the understanding of how tweets spread to different communities. A measure of *betweenness* in nodes connecting clusters we have identified here would likely be a strong starting point.

Our research highlighting the influence of communities, combined with related works detailing how these communities tend to be comprised of like-minded users illustrates that the conditions within Twitter allow for echo chambers to form and gain influence over the network. Echo chambers are thought to have played a significant role in the spread of "fake news" during the 2016 US election and our research validates this possibility. Applying our work to Twitter could help to identify areas of the network that facilitated the spread of misinformation and thereby provide insight on how to prevent it in the future.

Further studies into the network structure of Twitter are certainly warranted. As others test new clique detection methods and models of influence we believe that this thesis will prove a solid branching point into the ever expanding field of social network research.

References

[1] Eric Abrahamson and Charles J Fombrun. Macrocultures: Determinants and consequences. *Academy of Management Review*, 19(4):728–755, 1994.

[2] Eric Abrahamson and Lori Rosenkopf. Social network effects on the extent of innovation diffusion: A computer simulation. *Organization science*, 8(3):289–309, 1997.

[3] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

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

[5] applepie. tweepy. https://github.com/tweepy/tweepy, 2018.

[6] Richard P Bagozzi and Utpal M Dholakia. Intentional social action in virtual communities. *Journal of interactive marketing*, 16(2):2–21, 2002.

[7] Angela J Bahns, Christian S Crandall, Omri Gillath, and Kristopher J Preacher. Similarity in relationships as niche construction: Choice, stability, and influence within dyads in a free choice environment. *Journal of Personality and Social Psychology*, 112(2):329, 2017.

[8] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[9] W Lance Bennett and Jarol B Manheim. The one-step flow of communication. *The ANNALS of the American Academy of Political and Social Science*, 608(1):213–232, 2006.

[10] Immanuel M Bomze, Marco Budinich, Panos M Pardalos, and Marcello Pelillo. The maximum clique problem. In *Handbook of combinatorial optimization*, pages 1–74. Springer, 1999.

[11] Phillip Bonacich. Some unique properties of eigenvector centrality. *Social networks*, 29(4):555–564, 2007.

[12] Phillip Bonacich and Paulette Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social networks*, 23(3):191–201, 2001.

[13] Stephen P Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.

[14] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System sciences (hicss), 2010 43rd hawaii international conference on*, pages 1–10. IEEE, 2010.

[15] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177, 2001.

[16] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, P Krishna Gummadi, et al. Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30, 2010.

[17] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[18] Jere M Cohen. Sources of peer group homogeneity. *Sociology of education*, pages 227–241, 1977.

[19] Lisette De Vries, Sonja Gensler, and Peter SH Leeflang. Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of interactive marketing*, 26(2):83–91, 2012.

[20] Utpal Dholakia and Richard P Bagozzi. Consumer behavior in digital environments. *Digital marketing*, pages 163–200, 2001.

[21] Mario Diani. Social movements, networks and. *The Blackwell Encyclopedia of Sociology*, 2007.

[22] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[23] Susan T Ennett and Karl E Bauman. Peer group structure and adolescent cigarette smoking: A social network analysis. *Journal of health and social behavior*, pages 226–236, 1993.

[24] Jeffrey D Fisher. *Possible effects of reference group-based social influence on AIDS-risk behavior and AIDS-prevention.*, volume 43. American Psychological Association, 1988.

[25] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[26] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 211–220. ACM, 2009.

[27] Benjamin Golub and Matthew O Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49, 2010.

[28] Donald P Green, Bradley Palmquist, and Eric Schickler. *Partisan hearts and minds: Political parties and the social identities of voters*. Yale University Press, 2004.

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

[30] jeremylow. python-twitter. https://github.com/bear/python-twitter, 2018.

[31] Richard M Karp. Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Springer, 1972.

[32] Elihu Katz and F Paul. Lazarsfeld. 1955. personal influence: The part played by people in the flow of mass communications. *Glencoe, Illinois: The Free Press. KatzPersonal Influence: The Part Played by People in the Flow of Mass Communication*, 1955.

[33] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[34] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

[35] G Lawye. Understanding the spreading power of all nodes in a network: a continuous-time perspective. *arXiv preprint arXiv:1405.6707*, 2014.

[36] Paul Felix Lazarsfeld, Bernard Berelson, and Hazel Gaudet. The people's choice. 1944.

[37] Jingxuan Li, Wei Peng, Tao Li, Tong Sun, Qianmu Li, and Jian Xu. Social network user influence sense-making and dynamics prediction. *Expert Systems with Applications*, 41(11):5115–5124, 2014.

[38] Yung-Ming Li and Ya-Lin Shiu. A diffusion mechanism for social advertising over microblogs. *Decision Support Systems*, 54(1):9–22, 2012.

[39] Gabriele Lohmann, Daniel S Margulies, Annette Horstmann, Burkhard Pleger, Joeran Lepsien, Dirk Goldhahn, Haiko Schloegl, Michael Stumvoll, Arno Villringer, and Robert Turner. Eigenvector centrality mapping for analyzing connectivity patterns in fmri data of the human brain. *PloS one*, 5(4):e10232, 2010.

[40] Kazuhisa Makino and Takeaki Uno. New algorithms for enumerating all maximal cliques. In *Scandinavian Workshop on Algorithm Theory*, pages 260–272. Springer, 2004.

[41] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[42] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.

[43] Mezgrman. Tweetpony. https://github.com/Mezgrman/TweetPony, 2018.

[44] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498. ACM, 2014.

[45] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd international web science conference*, page 8. ACM, 2011.

[46] Pippa Norris. Did the media matter? agenda-setting, persuasion and mobilization effects in the british general election campaign. *British Politics*, 1(2):195–221, 2006.

[47] Pippa Norris and John Curtice. Getting the message out: A two-step model of the role of the internet in campaign communication flows during the 2005 british general election. *Journal of Information Technology & Politics*, 4(4):3–13, 2008.

[48] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814, 2005.

[49] John H Parmelee and Shannon L Bichard. *Politics and the Twitter revolution: How tweets influence the relationship between political leaders and the public*. Lexington Books, 2011.

[50] Wolfram Peiser. Cohort replacement and the downward trend in newspaper readership. *Newspaper Research Journal*, 21(2):11–22, 2000.

[51] Bohdan Pikas and Gabi Sorrentino. The effectiveness of online advertising: consumers perceptions of ads on facebook, twitter and youtube. *Journal of Applied Business and Economics*, 16(4):70–81, 2014.

[52] Dhavan V Shah and Dietram A Scheufele. Explicating opinion leadership: Nonpolitical dispositions, information consumption, and civic participation. *Political Communication*, 23(1):1–22, 2006.

[53] Tamara A Small. What the hashtag? a content analysis of canadian politics on twitter. *Information, communication & society*, 14(6):872–895, 2011.

[54] Twitter mau in the united states 2018 — statistic.

[55] Michael A Stelzner. Social media marketing industry report. *Social Media Examiner*, 41:1–10, 2011.

[56] Natalie Jomini Stroud. Media use and political predispositions: Revisiting the concept of selective exposure. *Political Behavior*, 30(3):341–366, 2008.

[57] Karthik Subbian and Prem Melville. Supervised rank aggregation for predicting influencers in twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 661–665. IEEE, 2011.

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

[59] Jimeng Sun and Jie Tang. A survey of models and algorithms for social influence analysis. In *Social network data analytics*, pages 177–214. Springer, 2011.

[60] Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363(1):28–42, 2006.

[61] Twitter. 2017 year end consolidated statement of operations. *Journal of Economic Perspectives*, 2017.

[62] Twitter usage statistics.

[63] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[64] Gabriel Weimann and Hans-Bernd Brosius. Is there a two-step flow of agenda-setting? *International Journal of Public Opinion Research*, 6(4):323–341, 1994.

[65] Debra A Williamson. Worldwide social network ad spending: a rising tide. *eMarketer. com*, 2:26, 2011.

[66] Shaozhi Ye and S Felix Wu. Measuring message propagation and social influence on twitter. com. In *International Conference on Social Informatics*, pages 216–231. Springer, 2010.

[67] Dan Zarrella. The science of retweets. *Retrieved December*, 15:2009, 2009.