TWITTER STRUCTURE AND FORMATION FROM INFORMATION PROPAGATION AND SECURITY PERSPECTIVE

by

MENG TONG

December 5th, 2013

A thesis submitted to the

Faculty of the Graduate School of

the University at Buffalo, State University of New York

in partial fulfillment of the requirements for the

degree of

Master of Science

Department of Computer Science and Engineering

Abstract

Online social networks like Twitter and Facebook are playing important roles in people's daily life nowadays. Due to their large user base, ubiquitous access through different classes of devices, easy production of content and high traffic, they are becoming ideal platforms for information propagation, both benign and malicious, with the latter leading to serious security issues. To better understand how information propagates on these networks, it is important to study their structure and formation process.

In this thesis, the structure and formation process of Twitter are studied to understand the issue relating to its role as a news media or another social network. It has been shown in the literature that the degree distribution of Twitter structure does not obey a power law distribution, as observed in many other online social networks. To address the issue relating to its role two large empirical datasets containing the whole topology of Twitter network have been analyzed, and the analyses suggests that Twitter structure has a component network following power law distribution, with the power law exponent similar to some other examples. Also, based on the analyses, we infer that different components of the Twitter network can be mapped to different roles in information propagation. The thesis then proposes a concise configurable model that can generate a network similar to the Twitter network in two steps. The model formation and its validity is verified by mathematical analysis as well as large scale simulation. The potential of extending this model to generate other online social networks and the impact as well as security implications of the proposed structure on information propagation has also been discussed.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Prof. Shambhu Upadhyaya for giving me such an interesting topic to work on, for his continuous support of my thesis work, for his patient guidance, encouragement and tolerence of my mistakes. His guidance helped me in all the time of research and writing of this thesis.

Besides my adviosr, I would like to thank Prof. Dimitrios Koutsonikolas for his support and encouragement, insightful comments and questions in every meeting we had.

My thanks also goes to Dr. Ameya Sanzgiri for his excellent work, which motivated my interest in this topic, for his help in discussing and inspiring my ideas, and using the tools.

Last but not least, I would like to thank my family and all my friends for the various support they gave me during this process.

Contents

List of Tables

List of Figures

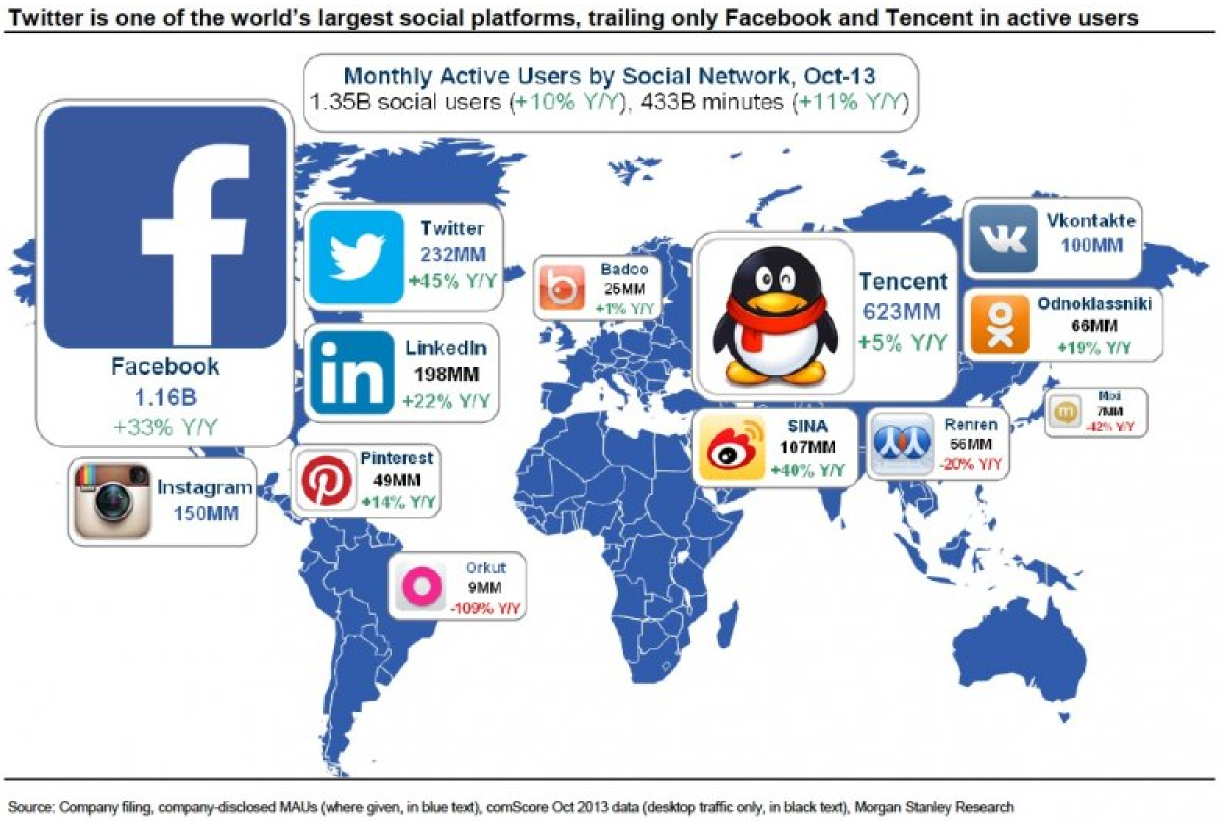
Chapter 1

Introduction

Online social networks are becoming an important part of our daily lives. Figure 1.1 shows a insta-snapshot of the world's biggest social networks: how many users they have, where they are dominant and how fast they are growing or shrinking [15]. There are different types of online social networks, including traditional social networks (e.g, Renren in China and Facebook elsewhere), microblog services (e.g, Sina Weibo in China and Twitter elsewhere), instant messengers (e.g, Tencent in China), job oriented networks (LinkedIn) and content sharing networks (Instagram). Due to various reasons, different social networks dominate different areas around the world.

Many of these online social networks share some of the following common characteristics, which make them an ideal platform for information propagation:

- **Large user base and fast growing speed**. Facebook already has more than one billion users. If it were a country, it will be the third largest in the world, after China and India. Many other online social networks also have hundreds of millions of users, and these numbers are still counting fast, especially for Twitter. This provides them


Twitter is one of the world's largest social platforms, trailing only Facebook and Tencent in active users

Source: Company filing, company-disclosed MAUs (where given, in blue text), comScore Oct 2013 data (desktop traffic only, in black text), Morgan Stanley Research

Figure 1.1: A map showing the battle between world's biggest social networks

with the ability to reach more people, compared with other methods of communication.

- **Ubiquitous access**. Today users can access these online social networks not only through websites, but also and maybe more frequently, from their mobile devices. Most of the online social networks provide mobile applications across different platforms, making their users always online. By carefully selecting the time to post a message, a user could target the audience with high probabilities [13].

- **Easy production of contents**. To produce and publish messages in online social networks is much easier, compared to other ways of communication. This gives ordinary people opportunities to get their voices and opinions heard, thus providing the online social networks a huge advantage over other traditional news media. On Twitter, a message (called a “Tweet”) is no more than 140 characters. Also, on other social networks users could simply publish a few words or a picture taken on the smart phones. This facilitates information propagation on these networks, however making it harder to control the quality of the contents.

- **High traffic**. There are large quantities of information going over these networks every second. Twitter states that there are more than 400 million Tweets per day [14]. The high traffic flow on these online social networks is a direct consequence of the previous three characteristics, and is also an important reason why these platforms are ideal for information propagation.

People have seen the power of these online social networks many times. For example, the miraculous landing of US Airways Flight 1549 on the Hudson River in New York in January 2009 was broadcast in real time via Twitter by onlookers who posted updates, reactions and even photos of the unfolding crisis, making it more dramatic[9]. Also, Twitter has become an important tool during disasters, as could be seen from the role it played during the 2011 Japan quake and tsunami. However, not all users are using Twitter in a proper way. In a recent event in April 2013, the Twitter account of the Associated Press was hacked and a bogus tweet was sent, causing Dow Jones Industrial Average falling more than 140 points within one minute. We can see how fast and far-reaching information propagation is on Twitter. Besides, it has been shown that Twitter suffers from the “click-hijacking” attacks[12].

In fact, all these are about information propagation. So, to address the security issues on these online social networks, it is of great importance to study the information propagation process. Studying the network structures is naturally the first step.

In this thesis an empirical study of the Twitter structure is conducted and the network formation models are proposed, as Twitter is one of the most powerful and important online social networks among these examples. The rest of the thesis is organized as follows: In Chapter 2 the background and related works, especially the knowledge of power law distribution, are introduced. Chapter 3 presents the empirical analysis of two large-scale Twitter datasets and the proposed models. The simulation results and discussions are included in Chapter 4. Finally, a summary of this thesis and a short discussion of future works are included in Chapter 5.

Chapter 2

Background and Related Work

2.1 Social Network Structure and Security Issues

Previous studies on online social network structures can be categorized into two different groups. One is empirical study by analyzing large-scale real world dataset and extracting their features[10] while others is to build models. The former methodology of research provides insights into the practical rules or behaviors that can be observed in many networks, like low diameters, high assortativities, while the other explains and predicts structures of networks[3]. In this thesis a combination of the two methods is tried by analyzing real world datasets first, and then building models based on the findings of empirical analysis.

Previous research on the information propagation and security could also be classified into two groups in a similar way as described above[17]. The first type of studies are empirical and to analyze information diffusion by large-scale datasets from existing online social networks. The second type of studies build mathematical models to analyze the mechanism by which information diffuses across the population[12]. The conclusion reached in this thesis could facilitate future research along both lines.

2.2 Power Law Distribution

Power law distribution has been observed in many situations of scientific interest across many disciplines from biology to economics. More importantly, it also plays a significant role in our understanding of complex network structures. This section provides the background of power law distribution that is necessary for further discussions.

2.2.1 Definition and Properties

Mathematically, the probability density function (PDF) of power law distribution can be defined as

$$p(x) = Cx^{-\gamma} \tag{2.1}$$

where C is a constant and γ is the scaling parameter of power law distribution. It has been shown that in real world datasets, scaling parameter γ typically falls in the range between 2 and 3, although with some exceptional cases[5]. Figure 2.1 shows a typical plot of PDF of power law distribution.

2.2.1.1 Scaling Parameter

As can be seen from the equation, scaling parameter governs the shape of the curve. Basically, a smaller scaling parameter will lead to a more even distribution than a larger one. This observation is very help in deriving our model in later chapters. See Figure 2.2 as an example of comparison between different scaling parameters. When the scaling parameter is equal to 1, the probability of larger value will be higher, and probability of smaller value will be lower, as compared with scaling parameters equal to 2 or 3.

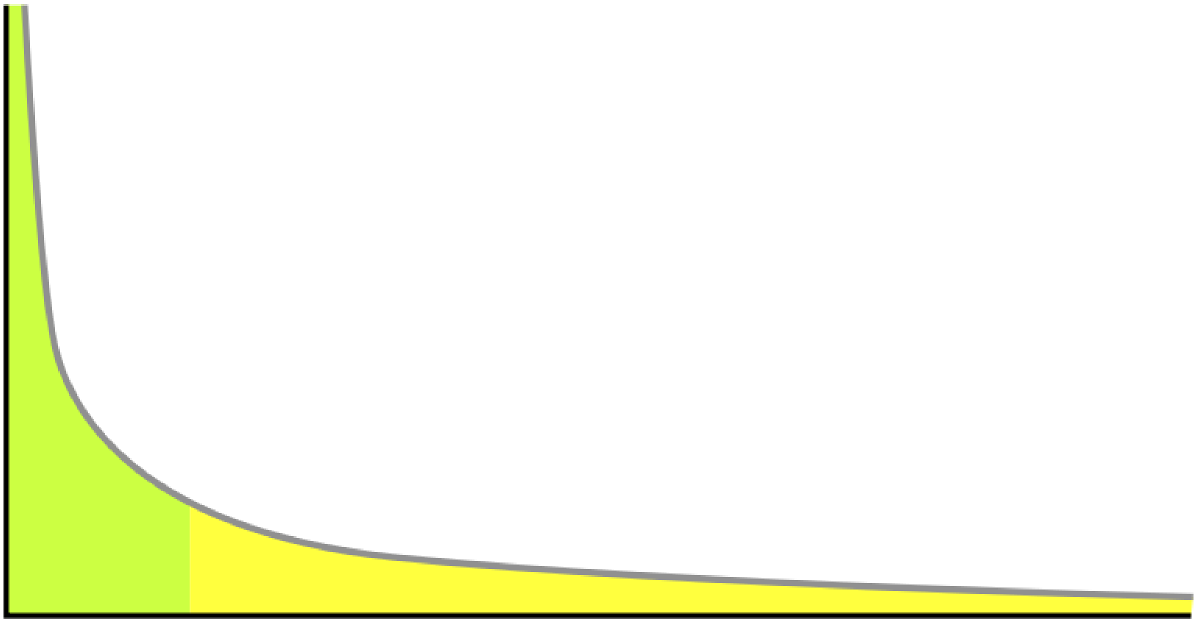


Figure 2.1: PDF Plot of Power Law Distribution[16]

2.2.1.2 Properties of Power Law Distribution

The most important property of power law distribution is the heterogeneity among the possible values. That is, a small number of values has high probabilities while the majority of values has very low probabilities of appearance. Thus unlike normal distribution, the average value cannot well describe variables that follow a power law distribution. This feature is observed in many economics phenomena and is the theoretical basis of the famous "long tail" theory and "80-20" rule. Given an undirected network with its degree distribution following a power law distribution, this means there exists a few nodes with a large number of connections, while the rest have relatively small number of connections.

2.2.2 Fitting of Power Law Distribution

An easy but not very accurate way to judge whether a dataset possibly follows a power law distribution is to plot its complementary cumulative distribution function (CCDF), on

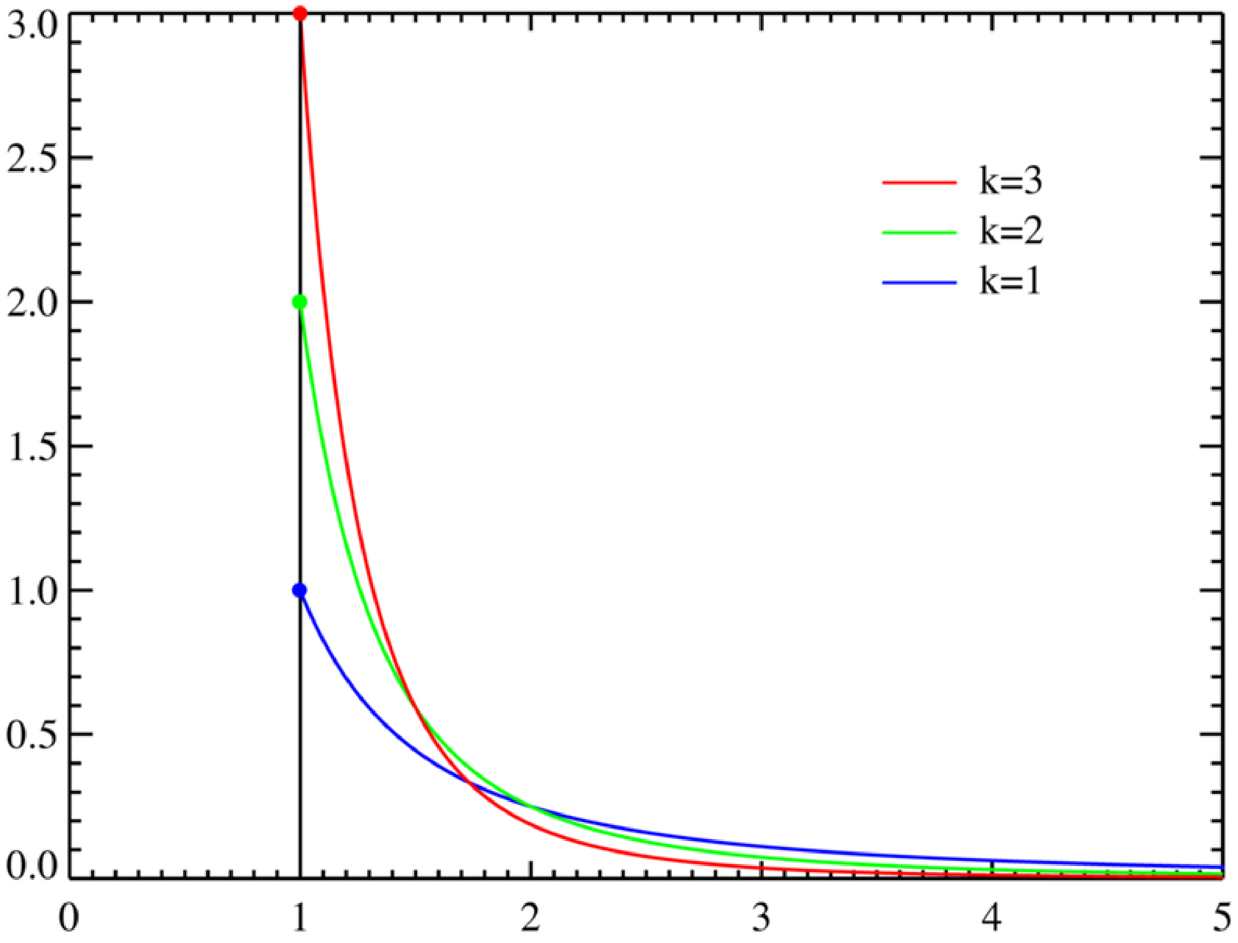


Figure 2.2: Power Law Distribution with Different Scaling Parameters[16]

a log-log scale, which will be a straight line:

$$\begin{aligned} P(X \geq x) =& \int_x^\infty Cx^{-\gamma} \\ =& \frac{C}{1-\gamma} x^{1-\gamma} \\ \log P(X \geq x) =& -C \log x \end{aligned}$$

This is also the first method used in analyzing the power law property of the empirical datasets.

Fitting power law distribution is non-trivial task, as pointed out in [5]. In our later analysis of an empirical dataset, fitting power law is a crucial step. Here the method described in [1] is adopted, which is an efficient implementation of the method described in [5].

2.2.3 Formation: Preferential Attachment

As many complex networks follow power law degree distribution, researchers have built formation models that could lead to such a degree distribution. The classical formation model is the preferential attachment model, first described in [3].

In this simple but elegant model, two features are crucial. First, the model assumes a growing network rather than forming links from a set of existing nodes. Secondly, when a new node joins, it has a larger possibility to connect to popular nodes than unpopular nodes, or in other words, it prefers to connect or attach to existing nodes with higher degrees, thus the name “preferential attachment.”

Let m be the number of links a new node forms upon joining, $d_i(t)$ be the degree of node i

at time t, then when a new node joins, a current existing node i gains

$$m \frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)} \tag{2.2}$$

new links. Note that mean field approximation approach is used here by assuming that every new node forms the same number of links. When different behaviors of different nodes are considered, it become almost impossible to model the process. Thus instead the average behavior is considered, and has been proved to be a good approximation[6].

This gives the increasing rate of $d_i(t)$ to be

$$\frac{dd_i(t)}{dt} = m \frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)} \tag{2.3}$$

where $d_i(t)$ is the degree of node i at time t. Solving this differential equation with a start condition of $d_i(i) = m$ will give a power law degree distribution with a scaling parameter equal to 3. The detailed process will be illustrated in the derivation of our proposed model in Section 3.5.2.

Though the original preferential attachment works on undirected network and resulting in a power law distribution with scaling parameter equal to 3, it is easy to modify the model to incorporate directed networks and yielding a wide range of the scaling parameter. This is also shown in the proposed model in Section 3.5.2.

Chapter 3

Empirical Study and Proposed Model

3.1 Online Social Networks Features

Many online social network structures have been extensively studied, and they are shown to adopt some of the following common characteristics, compared with other random networks[7]:

- The average distance between pairs of nodes in a social network is small, which leads to the famous "six-degree of separation."
- The clustering coefficient is large.
- The degree distribution follows power law, so that connections are highly concentrated to a few number of nodes.
- Assortativity: nodes tend to connect to other nodes with similar degree.
- Inverse clustering: neighbors of a high degree node are less likely to be linked to each other.

While the combined effect of all these features actually effect the information propagation property and security characteristic of a certain network, it's hard to study them separately and put them together. Instead, a more effective way is to start from degree distribution, the fundamental property of a social network, which could thus lead to some of the other features.

As has been shown in [10], many online social networks are likely to have a degree distribution following a power law distribution, with the scaling parameter falling in the range between 2 and 3. However, Twitter[8] and Facebook[2], the two most famous and most recognized online social networks, are exceptions, as their degree distributions do not obey the above rule. Motivated by [12] and other studies on Twitter, it becomes interesting for us to explore the structure of Twitter by first analyzing some empirical datasets.

3.2 Dataset Information

The Twitter dataset from [8] and [4] are used in this thesis, and is denoted as $D1$ and $D2$, respectively. Table 3.1 lists the basic information from the two datasets.

Dataset	Users	Links	Time Obtained
$D1$	41,652,230	1,468,365,182	Jun. 2009
$D2$	52,579,682	1,963,263,821	Sep. 2009

Table 3.1: Datasets Basic Information

We can see the fast growing speed from here that the number of Twitter users increased more than 25% in just three months. But more importantly, it should be noted that these two datasets contain the whole topology of Twitter network at the time they were crawled. This provides the "truth" of Twitter network, rather than a sampled dataset or a part of it. Such a large network is shaped by all its users, thus any conclusions reached from a partial network is not convincing enough. Besides, by studying the difference of the two datasets crawled in a close time period, the evolution trend of Twitter network would be more clear. Based on the above reasons, we use these two datasets in our empirical study.

3.3 Empirical Study

In [8] the authors raise two points that are particularly interesting to us. The first is whether Twitter is a social network or a news media. The second is the conclusion of Twitter having a non-power-law follower distribution.

The first question is hard to answer. The main reason is that this is largely user-dependent, which means different users use Twitter as different tools. Some people use Twitter as a news media or information resource, by following some public accounts like BBC or Fox news while some other users mainly use Twitter to interact with their friends in the real world, by following mostly their friends in the real world, or the "off-line" social network. In this way they could get whatever news tweeted by their friends immediately. And of course there exists users who use Twitter in both ways. Part of their focus is on public news and part of the attention would be drawn to the news produced by their friends. This would make their timelines rather busy.

Despite the complicated behavior of the users, some differences between these two different usages could be discovered. Specifically, if a user follows another user as a source of information, typically this would be a one way relationship. For example, if Alice decides to follow "ESPN" for its hot sports news, then normally everything stops there when she clicks the "Follow" button, since "ESPN" doesn't know Alice and hence will not follow her back. It is very much the same situation when people follow other celebrities such as famous actors, sport stars or organizations. However, on the other hand, when a user follows his or her friend in the real world, there is a good chance that this friend will follow back. For example, if Alice and Bob are members of a same club and when Alice opens a Twitter account and follows Bob who is already on Twitter, it is likely that Bob will also follow Alice, typically in a very short time.

3.3.1 Network Separation

The discussion above suggests that the analysis of the Twitter network needs to be from a different perspective. Instead of a single network, why cannot it be separated into two different networks, according to their different purposes? The two subnetworks so extracted from the Twitter network could be defined as follows:

- **Social Network**: a network containing all mutual relationships. This is an undirected network where every pair of connections implies that the connected users mutually follow each other on Twitter.

- **Information Network**: a network containing all the one way relationships. This is a directed network where every pair of connections implies that one user follows the other but NOT vice versa.

Figure 3.1 illustrates this network separation. There are three different types of nodes under this network separation:

1. Nodes only in information network. These correspond to the yellow nodes in information network in 3.1. They have only one way relationships.

2. Nodes only in social network. These correspond to the blue node in social network that

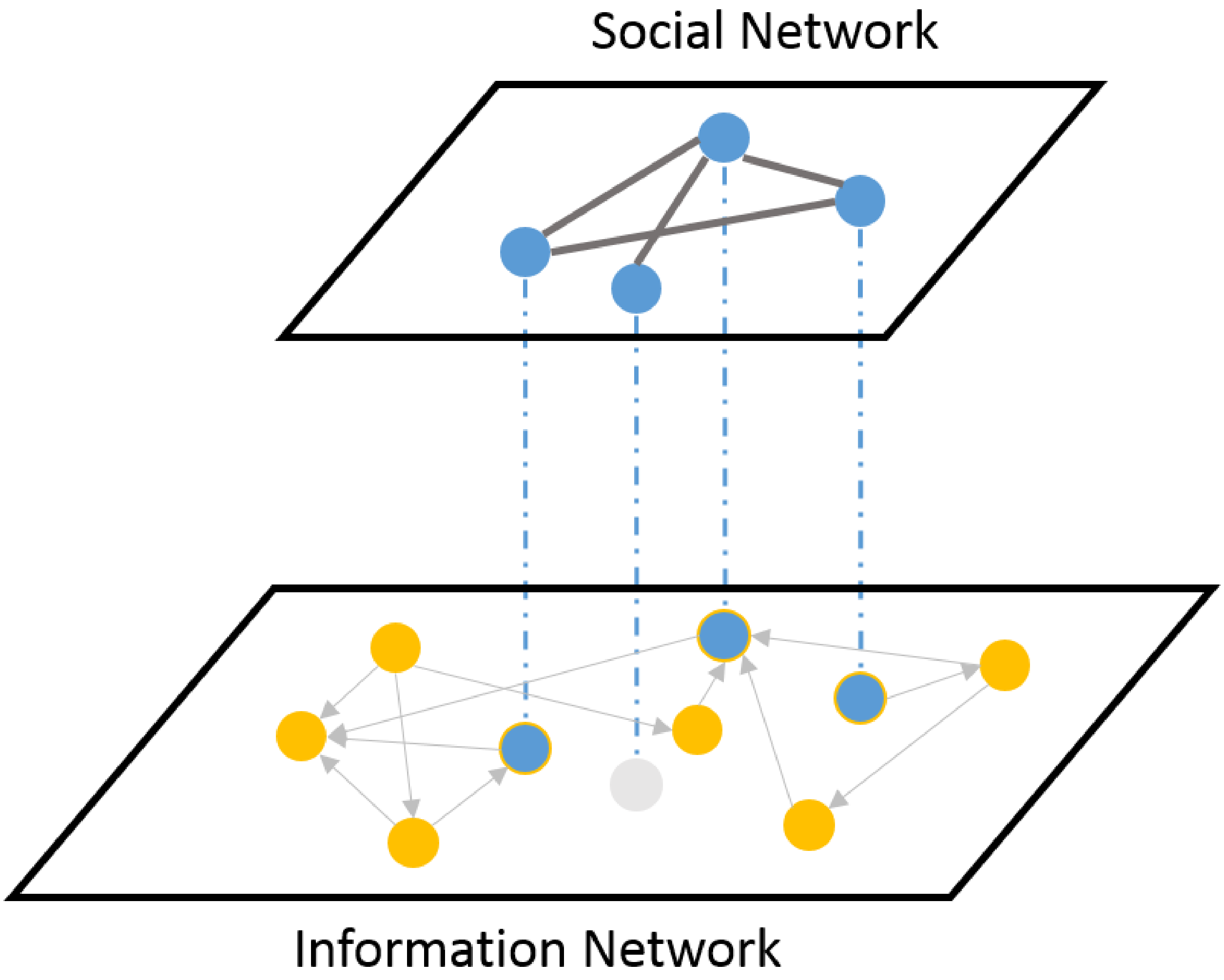


Figure 3.1: Network Separation. Top: Social Network. Bottom: Information Network. Blue nodes appear in both networks. Yellow Nodes appear only in information network. Links in information network are directed and lighter, indicating its lower level of trust. Links in social network are undirected and darker, indicating its higher level of trust.

has a gray projection (indicating not existing) in information network in 3.1. They have only mutual relationships.

3. Nodes in both networks. These correspond to the blue nodes in social network that has a blue projection in information network with a yellow edge in 3.1. They have both mutual relationships and one way relationships. It should be noted that if two nodes are connected in social network and both appear in information network, they will not be connected in information network.

These three types of nodes correspond to the three different types of users discussed in previous sections. The nodes in these two networks could be overlapping but the links in these two networks are mutually exclusive. That is, a link cannot appear in both networks.

It should be noted that here only the follower network is considered, that is, the out degrees of all the nodes are considered. The reason is that one user's tweets will appear in all his or her followers' timeline, but has no effect to his or her friends' timelines. From the perspective of information propagation, the out degree of a node indicates how many other nodes it could reach in one hop, that is its ability to spread information from the point of view of size. So it is meaningful to study the follower network, rather than the friend network.

In order to conduct our theoretical analysis, the overall Twitter follower network is denoted as a graph $G_a = (V_a, E_a)$. Here V_a is a set containing all the nodes appearing in Twitter follower network, $V_a = \{v | d(v) \geq 0\}$. As users with no followers have no ability to spread information, it is assumed that all $v \in V_a$ have a non-negative number of followers. E_a is a set containing all the follower relationships in the network. If $e_{ij} \in E_a$, then user i follows user j.

We define the social network as $G_s = (V_s, E_s)$ and information network as $G_i = (V_i, E_i)$. Then we have the following relationships:

$$V_s \cup V_i = V_a$$
$$E_s \cup E_i = E_a \quad E_s \cap E_i = \emptyset$$
$$e_{ij} \in E_s \Rightarrow e_{ji} \in E_s$$

This separation is also meaningful in the following sense:

- Level of trust. People treat friends differently from strangers in real world in that they trust friends more. This is also true in online social networks. It has been shown in [12] that in the form of "Click-jacking" attack, the probability of a user clicking a link embedded in a tweet has a large effect on the overall success of this attack. However this value is not the same across all the friends of a user, it is reasonable to assume a higher level of trust in the social network and lower level of trust in the information network, though the nodes in information may be some celebrities. This will lead to a different clicking probabilities of links and thus provides a better understanding of the security issues.

- Formation process. These two networks should be formed in different ways. It is reasonable to assume that when selecting information sources, people first consider popular nodes, as popularity is linked with their authority. But when choosing friends, it is hard to tell what is the underlying principle governing the selection behavior. Different processes have been tried in building our model. From this point of view, it is also meaningful to separate the social network from the information network.

Then based on this network separation, it is hypothesized that the two subnetworks, social network and information network, should have more clear degree distribution. Specifically, it could be the case that one or both of them would fit a power law distribution very well. And since the original Twitter follower network could be regarded as a mixed-purpose network or a combination of social network and information network, it is reasonable that it does not strictly follow a power law distribution. The testing of our hypothesis is described in the next section.

3.4 Fitting and Results

To test our hypothesis on the empirical datasets, the social network and the information network are first extracted from the originally combined datasets. This is not a trivial task due to the large sizes of the datasets ($27GB$ and $37GB$, respectively). A divide-and-conquer strategy is adopted to process the dataset piece by piece. Table 3.2 shows some statistics about the extracted social network considering only nodes with mutual followers.

Attribute	$D1$	$D2$
Total users	41,652,230	52,579,682
Users in social network	22,580,393	26,866,589
Average degree in social network	23	25
Maximum degree in social network	698,112	713,207

Table 3.2: Social Network Statistics

To our surprise in both datasets there are only about 50% of Twitter users are following at least one other user that is also following itself. This corroborates the conclusion reached in [8] that Twitter has a lower level of reciprocity of 22% in $D1$ (21.6% in $D2$), compared with other online social networks like Flicker and Yahoo! 360. The huge difference between the maximum degree and the average degree in social network also suggests the heterogeneity among the nodes, which could possibly lead to a power law distribution.

Attribute	$D1$	$D2$
Total users	41,652,230	52,579,682
Users in information network	38,355,089	47,175,611
Average degree in information network	24	26
Maximum degree in information network	2,997,304	3,503,476

Table 3.3: Information Network Statistics

Table 3.3 shows the statistics for information network. Only a small fraction of users of Twitter has no followers. Also, the difference between the average degree and maximum degree is significant.

It should be noted that although $D2$ is larger in size than $D1$, they show the same basic properties in the two separated networks. This indicates that Twitter may have stepped into a stable stage in its revolution.

The next step is to check whether the degree distribution of these two subnetworks follows power law distribution. To do this the complementary cumulative distribution function (CCDF) in normal scale and log-log scale is plotted first. As mentioned in 2.2.2, fat tail feature is expected to be observed in the normal scale CCDF plot, and a straight line in the log-log scale plot. Figures 3.2 and 3.3 show the plot of $D1$. Since the plots for $D2$ is almost the same, they are not shown here.

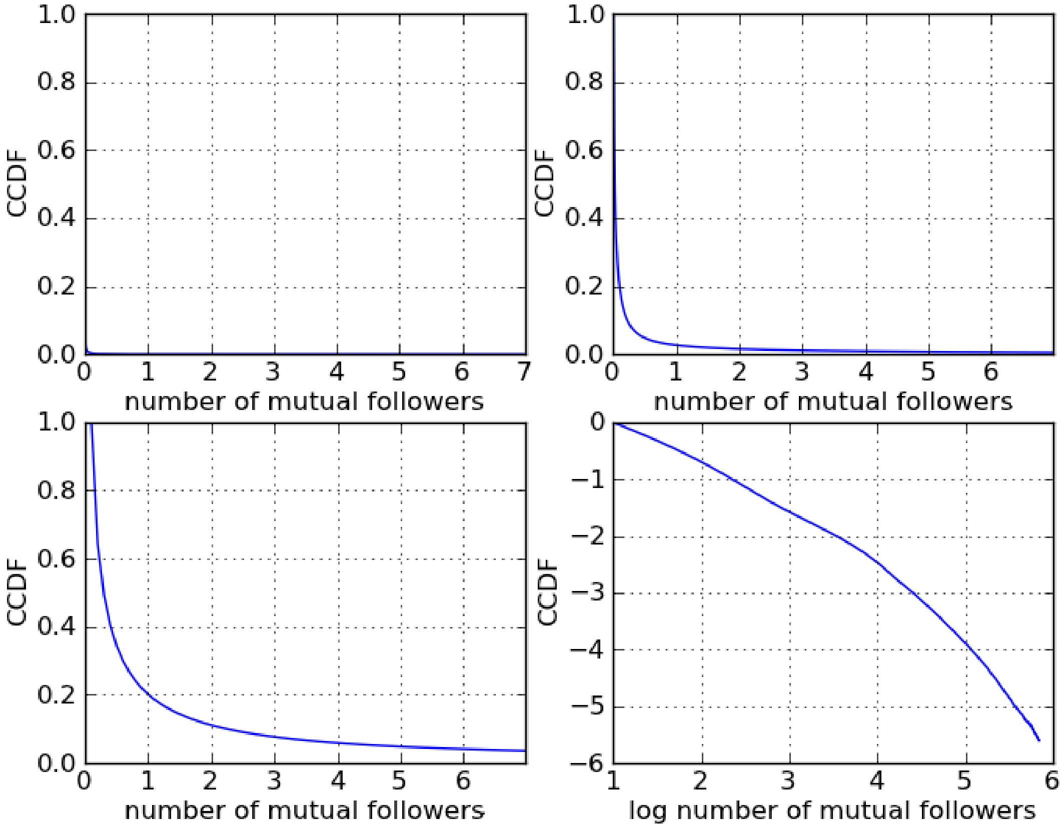


Figure 3.2: CCDF of Social Network Degree in $D1$. Top Left: CCDF in normal scale. Top Right and Bottom Left: Zoomed versions of CCDF in normal scale (notice the change of scale in x-axis). Bottom Right: CCDF in log-log scale

The plotting result is very encouraging since both log-log plots are very close to a straight line, starting from a lower bound. This indicates that our hypothesis has a high probability to be true. However, we cannot conclude with certainty of the power law rule simply by the plot. Moreover, in addition to confirm these power law distributions, the exact scaling parameters are also needed. As described in [1], the dataset is fit into power law distribu-

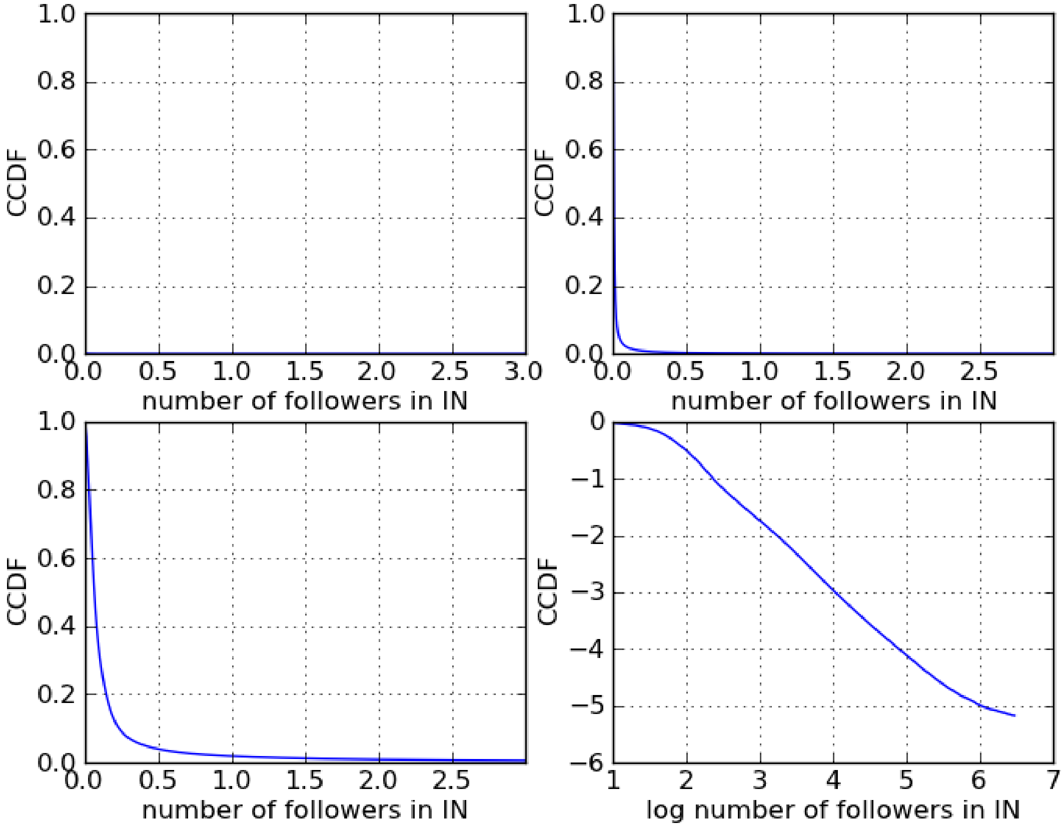


Figure 3.3: CCDF of Information Network Degree in $D1$. Top Left: CCDF in normal scale. Top Right and Bottom Left: Zoomed versions of CCDF in normal scale (notice the change of scale in x-axis). Bottom Right: CCDF in log-log scale

tion, and the scaling parameter as well as the goodness of fit compared with other candidate distributions are calculated. The fitting results are shown in Table 3.4. The comparison with alternative distribution is shown in likelihood ratio.

Network	Scaling Parameter	Exponential	Lognormal
Social Network in $D1$	1.87	293	-18
Social Network in $D2$	1.88	309	-24
Information Network $D1$	2.24	34	28
Information Network $D2$	2.15	155	10.7

Table 3.4: Power Law Fitting of Social Network and Information Network

Basically the information network is a good fit of power law distribution with scaling parameter equal to 2.24 in $D1$ and 2.15 in $D2$, compared with exponential lognormal distribution. However, for social network the power law fitting does not have big advantage over lognormal distribution. In fact, as the fitting algorithm is not deterministic, sometimes power law is a better fit the lognormal but sometimes not. In our proposed model in later sections, different models are tried for these two possibilities.

3.5 Generation of Proposed Models

In this section a series of two-step configurable formation models is proposed in order to generate a similar network capturing the degree distribution of the real Twitter network.

It is important to point out that the goal here is not simply to generate a network with its degree following power law distribution. Instead, the aim is to find a process that is similar to Twitter user behavior as much as possible and could lead to a similar network distribution at the same time.

3.5.1 Real World Process

Before building the models it is helpful to take a look at what Twitter suggests its users to do upon opening a Twitter account. This is divided into three steps:

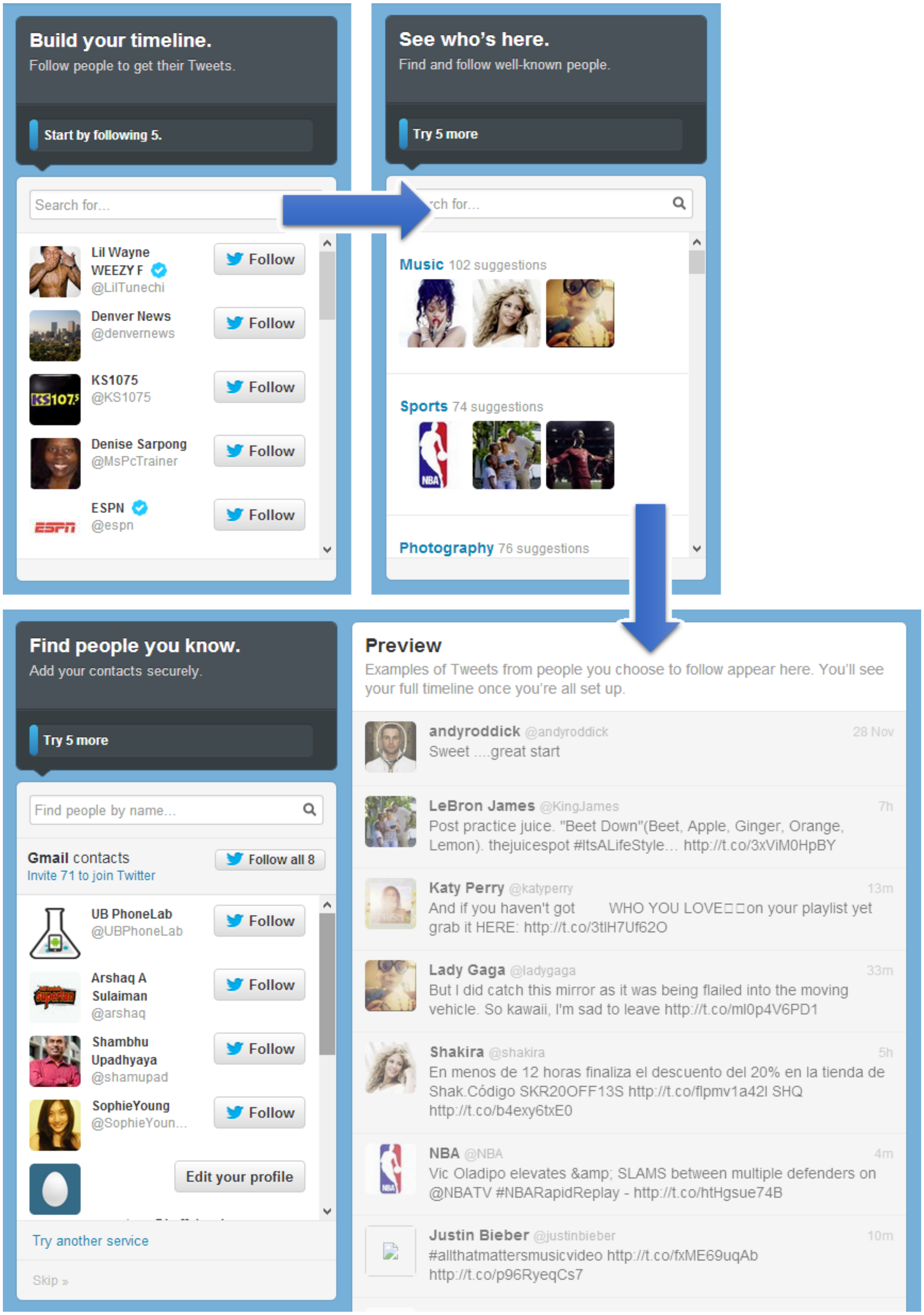
1. When you open a new Twitter account, Twitter will first provide a list of popular users (For example, celebrities like Ashton Kutcher, news media like ESPN) and ask you to follow 5 of them (do not need to be the suggested ones if you search by name).

2. Twitter will analyze the areas of interest based on your selections in the first step. Then it will provide another categorized list and ask you to follow 5 more (again, do not need to be the suggested ones). As 2 singers and 2 sports related organizations NBA and ESPN are selected in the first step in this example case in Figure 3.4, the provided categories are music and sports.

3. In this step, Twitter will ask for permission to access your contact list on gmail or yahoo! mail to find people you know that are using Twitter and suggest you to follow 5 of them.

The three steps are briefly illustrated in Figure 3.4.

This is a very interesting process because it almost clearly (by almost it means still some of your friends will appear in the provided lists in step 1 and 2) separates the formation of the two subnetworks, information network and social network. The first two steps can be regarded as helping the new user forming the information network, by suggesting popular



Build your timeline.
Follow people to get their Tweets.

Start by following 5.

Lil Wayne WEEZY F @LilTunechi

Denver News @denvernews

KS1075 @KS1075

Denise Sarpong @MsPcTrainer

ESPN @espn

See who's here.
Find and follow well-known people.

Try 5 more

Music 102 suggestions

Sports 74 suggestions

Photography 76 suggestions

Find people you know.
Add your contacts securely.

Try 5 more

Gmail contacts
Invite 71 to join Twitter

UB PhoneLab @UBPhoneLab

Arshaq A Sulaiman @arshaq

Shambhu Upadhyaya @shamupad

SophieYoung @SophieYoun…

Try another service

Skip »

Preview
Examples of Tweets from people you choose to follow appear here. You'll see your full timeline once you're all set up.

andyroddick @andyroddick 28 Nov
Sweet ....great start

LeBron James @KingJames 7h
Post practice juice. "Beet Down"(Beet, Apple, Ginger, Orange, Lemon). thejuicespot #ItsALifeStyle… http://t.co/3xViM0HpBY

Katy Perry @katyperry 13m
And if you haven't got WHO YOU LOVE□□on your playlist yet grab it HERE: http://t.co/3tlH7Uf62O

Lady Gaga @ladygaga 33m
But I did catch this mirror as it was being flailed into the moving vehicle. So kawaii, I'm sad to leave http://t.co/ml0p4V6PD1

Shakira @shakira 5h
En menos de 12 horas finaliza el descuento del 20% en la tienda de Shak.Código SKR20OFF13S http://t.co/flpmv1a42l SHQ http://t.co/b4exy6txE0

NBA @NBA 4m
Vic Oladipo elevates &amp; SLAMS between multiple defenders on @NBATV #NBARapidReplay - http://t.co/htHgsue74B

Justin Bieber @justinbieber 10m
#allthatmattersmusicvideo http://t.co/fxME69uqAb http://t.co/p96RyeqCs7

Figure 3.4: What happens when opening a new Twitter account. Top Left: Step 1. Top Right: Step 2. Bottom: Step 3.

existing users in the fields attracting the new user. The third step is obviously building the social network, by importing from other existing social relationships, typically people that in the email contact list that are using Twitter. This is very reasonable because for information sources since you always want something with public trust, thus the popular existing users would be good choices because they are trusted by a large amount of users. This is a basic characteristic of the preferential attachment model. Twitter also prompts with "Find and follow well-known people" in the second step. However when building the social network it is largely based on your "social network" in the real world, which could be somehow reflected by the contact list of email accounts.

This process also reveals two other important numbers in building the model. The first is the total number of users that a user will follow upon joining Twitter, which is 15 if the user strictly follows the Twitter suggestions. The second one is the ratio between the number of users a new user follows by searching his or her contact list and the total number of users the new user follows, denoted as α. Parameter α is called "social ratio" in this thesis and this name will be explained later. Again if the user strictly follows the Twitter suggestions, α will be 1/3, as 5 out of 15 of the users newly followed are supposed to be "real friends."

3.5.2 Description of Proposed Models

Based on the observations from the real world process and the above analysis, a series of two step configurable models are proposed.

These models are described as follows. At each time step, there is a new node joining the network, making it a growing network formation model. Upon joining the network, the new node selects m existing nodes with whom to form a relationship. Among these m nodes, some of them are selected as information source, while others are selected because they are friends in the real world. Nodes selected as information source appear in the information network, by forming a directed link from the new node to them, which is named as "information network nodes", and the links formed in this way are called "information link" in the following context. Nodes selected because of real-world friendship appear in the social network, by forming two directed links from the new node to them and in the reverse direction also. They are called "social network nodes" or "mutual followers", and the links formed in this way are called "social links", in the rest part of the thesis depending on the relationship established. Assume that all the mutual following relationships are formed when a new node joins the network, which is not realistic but would greatly simplify the model. As previously defined, let α be the "social ratio" representing the ratio between social network nodes and the total number of users followed by this new user. Thus there are $(1-\alpha)m$ information network nodes and αm social network nodes.

It takes the new node two steps to select all m nodes to connect with (it will be shown in simulation that the order of the two steps could be changed).

1. **Select information network nodes**. This is based on the preferential attachment scheme, that is, the probability of an existing node to be selected as an information network node by the new user is proportional to its current in-degree. This is similar to Twitter suggesting users to follow popular users. However it should be noted that the current in-degree of an existing node includes not only its information network in-degree, but also its social network in-degree. The reason could be stated from two sides. From the existing node's point of view, the social network in-degree also contributes to its popularity. From the new node's point of view, when it is considering whether or not to follow another user as information source, it will not separate the existing

node's mutual followers with pure followers.

2. **Select social network nodes**. As has been mentioned, the principles of people selecting social network nodes are largely dependent on their real world social networks. Thus it is relatively difficult to model this process within the Twitter environment. And this may also be the reason why Facebook network distribution cannot be explained by a power law fitting. As the empirical fitting result shows that the power law fitting of social network degree distribution has no big advantage over lognormal distribution, two different processes are tried to determine their effect.

    - Preferential attachment. This means in selecting the social network nodes, the new node will also connect to popular nodes. However, here only the social network in-degree is considered. This is referred to as model I in the following context.
    - Multiplicative process. In [11] it has been shown that a multiplicative process will generate a lognormal distribution. This process is simulated by randomly selecting social nodes from the set of existing nodes. This is referred to as model II in the following context.

    It should also be noticed that the formation process of social network is independent from the formation of information network, but not vice versa. The effect of these two options for social network nodes selection is tested in the simulations.

In the real world scenario, the values of m and α are different across all the users. However it would be too complicated to model those differences at the same time. Instead, the mean field approximation is used in these models by assuming that every user behaves like an average user with the same nodes to connect to, and with the same friends to select. It will be shown in simulation that the behavior of "social ratio" α may only be effected by its expectation.

These models are configurable in that the “social ratio” α could be adjusted. A brief discussion on the possibility of α as a intrinsic property of a online social network is provided after the simulation results.

3.5.3 Mathematical Analysis

In this section the proposed models are analyzed mathematically.

Let $d_k^i(t)$ be the in-degree of node k in information network at time t, $d_k^s(t)$ be the in-degree of node k in social network at time t, and $d_k(t)$ be the total in-degree of node k at time t. When a new node joins at time t, the number of new information links an existing node k will gain is

$$(1-\alpha)m\frac{d_k^i(t)+d_k^s(t)}{\sum_{j=i}^{t} d_j(t)} \tag{3.1}$$

Similarly, under preferential attachment selection of social nodes, the number of new social links an existing node k will gain is

$$\alpha m\frac{d_k^s(t)}{\sum_{j=i} d_j^s(t)} \tag{3.2}$$

Or, under random selection of social nodes, the number of new social links an existing node k will gain is:

$$\frac{\alpha m}{t} \tag{3.3}$$

Solving the equations for social network is easy because they do not depend on other parameters. Equation 3.2 gives solution

$$d_k^s(t) = \alpha m(\frac{t}{k})^{0.5} \tag{3.4}$$

Substitute 3.4 back to 3.1 gives

$$\frac{dd_k^i(t)}{dt} = \frac{d_k^i(t)}{At} + \frac{\alpha m}{Ak^{0.5}} \times \frac{1}{t^{0.5}} \tag{3.5}$$

where $A = \frac{1+\alpha}{1-\alpha}$ Solving this differential equation will give

$$d_k^i(t) = \frac{2\alpha m}{2-A} \times ((\frac{t}{k})^{\frac{1}{A}} - (\frac{t}{k})^{\frac{1}{2}}) \tag{3.6}$$

Compared with Equation 3.4 it could be noticed that there are two power law components here. So the resulting scaling parameter here is effected by α and the network structure of social network part. A larger α will lead to a larger scaling parameter of information network.

Chapter 4

Simulation and Results

4.1 Simulation Setup

All simulations start with an initial network containing m_0 nodes, fully connected with each other. The reason to let the initial network be fully connected is to mimic the launching process of Twitter, as well as many other online products. Basically they would start with invitation or internal test, which indicates highly connected relationships between the initial users.

As described in 3.5.2, the formation of the network continues by adding one node at each time step. Upon joining the network the new node would select information nodes and social nodes, and the network will get updated. If a node is selected as the one of the two types, then it will not be selected again.

In summary, the effect of the following parameters on the fitting result will be simulated on the proposed models:

- Social ratio: α
- Initial network size: m_0
- Number of nodes to form a relationship with when a new node joins: m

All simulations are stopped when the network size reaches 0.6 million, which is large enough as can be shown from the results below.

4.2 Effect of "social ratio" α

4.2.1 Fixed α

Figure 4.1 and 4.2 show the effect of "social ratio" α along the way of the network evolution, for the two models proposed in Section 3.5.2. In these simulations, m and m_0 are both set to be 20. Different values of α are tested on both models and the values are selected so that αm and $(1-\alpha)m$ are both integers.

Although there are glitches in the curves, in general, all of them show the same trend for scaling parameter γ: γ increases in the beginning of the network evolution, and eventually saturates at a stable value γ_s. The saturation scaling parameter γ_s is our focus since it occurs when the network size approaches infinity. Figure 4.3 shows the saturation scaling parameter α for different situations.

It can also be observed from the figures that larger α produces larger scaling parameter γ, which is consistent with the mathematical deduction.

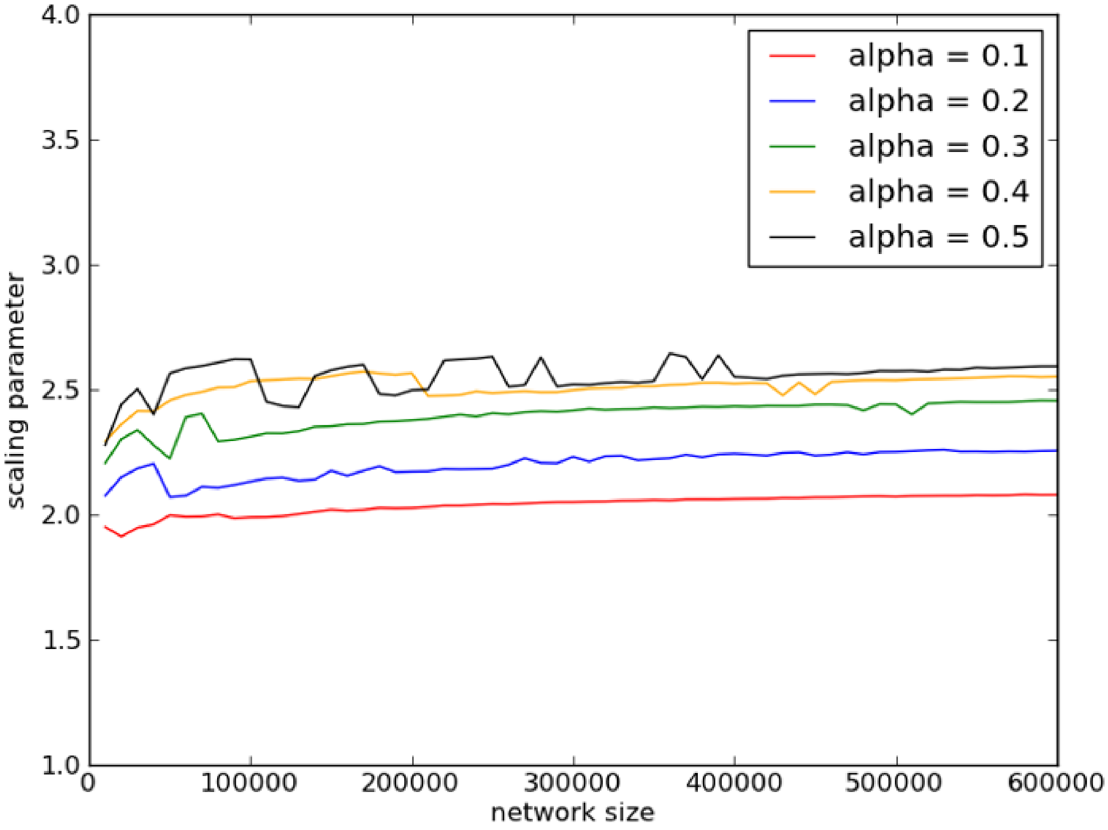


Figure 4.1: Effect of “social ratio” α on the result scaling parameter during evolution of the network (Model I) X-axis represents the network size. Y-axis represents the fitting result of the scaling parameter α of the information network. Different curves represent the fitting results for different social ratio α. (Other parameters: $m = 20$, $m_0 = 20$).

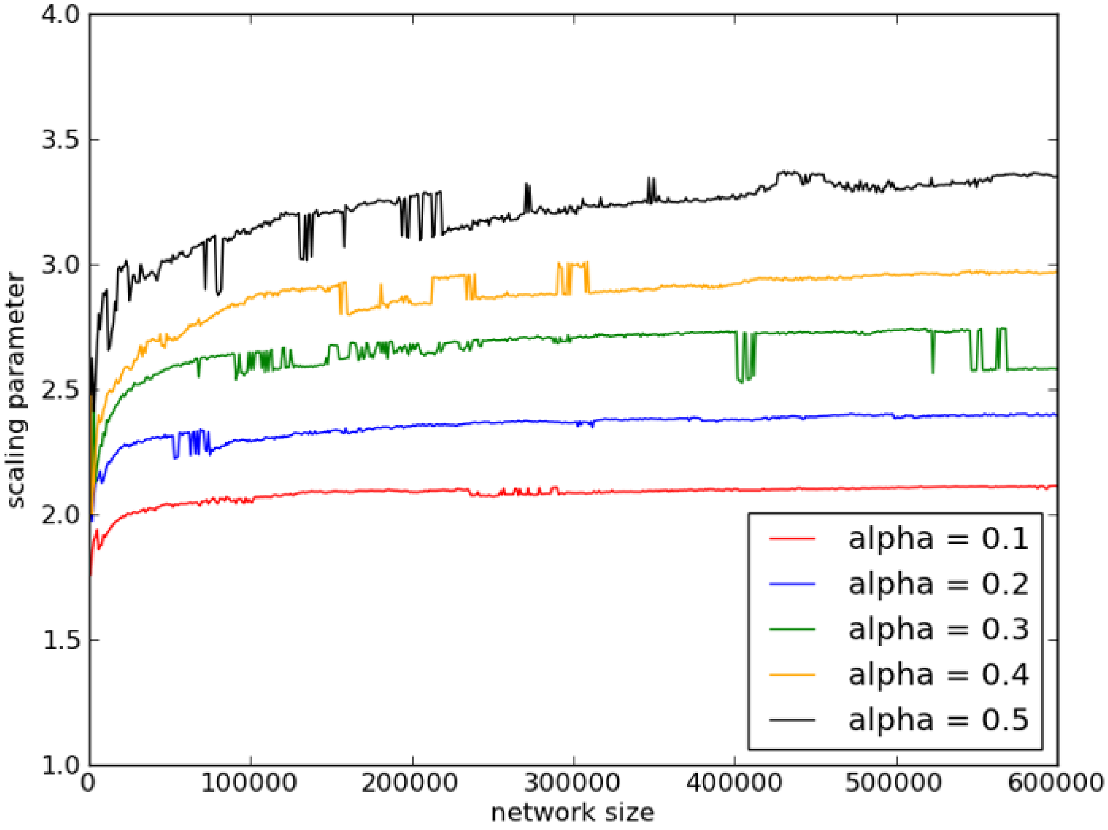


Figure 4.2: Effect of “social ratio” α on the result scaling parameter during evolution of the network (Model II) X-axis represents the network size. Y-axis represents the fitting result of the scaling parameter α of the information network. Different curves represent the fitting results for different social ratio α. (Other parameters: $m = 20$, $m_0 = 20$).

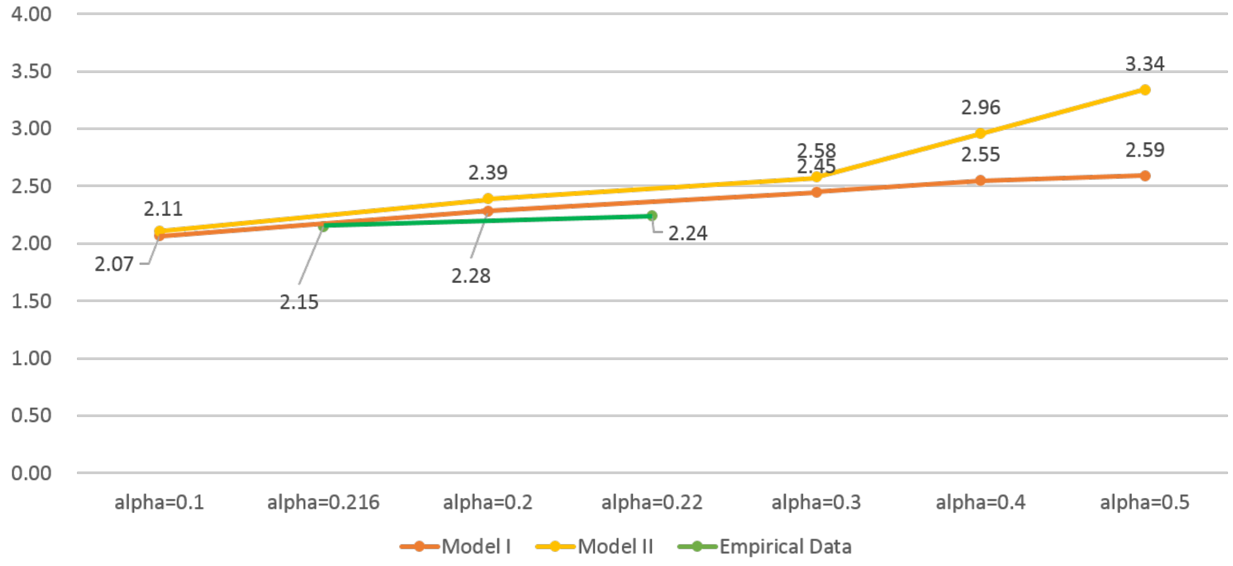


Figure 4.3: Saturation γ_s for Different Models and α. Social ratios equal to 0.216 and 0.22 are fitted from empirical dataset. They are not tested in simulations.

To compare the two different models, closer attention should be paid to the two blue lines that representing "social ratio" α equal to 2 since they are closest to the overall α in the empirical dataset, which is reported to be 0.22 in[8]. In Figure 4.3, the saturation scaling parameter γ_s is 2.28 for model I and 2.39 for model II. Since the empirical information network has a scaling parameter of 2.24 when $\alpha = 0.22$ and 2.15 when $\alpha = 0.216$, it can be concluded that Model I is an enough good fit for generating the desired information network.

However, the fitting of the social network part remains an open question. As analyzed in Section 3.5.3, the social network structure has an impact on the result fitting scaling parameter of the information network while itself is independently formed from the information network. Since changing the selection of social network nodes from random selection to preferential attachment decreases the saturation scaling parameter γ_s from 2.39 to 2.28, it is reasonable to make the hypothesis that a social network with a lower scaling parameter power law distribution will further decrease the saturation scaling parameter γ_s a little bit, yielding a value even closer to 2.24 or 2.15 However, the preferential attachment scheme cannot produce an undirected network with scaling parameter around 2, suggesting that human factors outside the scope of degree should be brought into consideration.

It is interesting to think about the role that the “social ratio” α played in the network formation process and resulting structure. Generally, it reflects on average how social are the users in this network. Here how “social” could be interpreted as how often or how willingly are the users of this network to use it as a social network with their friends. Thus, it could be regarded as an intrinsic property of a particular network, determined by the nature of the network. In these two empirical dataset, it is calculated that the social ratio α is decreasing, from 0.22 in Jun.2009 to 0.216 in Sep.2009, indicating that users are getting more and more information-driven on Twitter. From this point of view, it is meaningful to extend this model to general networks that have both mutual and one-way relationships.

Based on the comparison between model I and II, the following simulations will mainly be focused on model I, since it is better of the two models.

4.2.2 Different α for Different Users

It is assumed in the models that all users will have the same “social ratio” α, under the mean field approximation approach. The goodness of this approximation is tested by assigning to α with a distribution similar to that calculated in the empirical dataset, as shown in Figure 4.4. It could be observed from the figure that a large fraction of the users has no or very small percentage of mutual friends, and some values of α are more frequent than others. This distribution is simulated by randomly picking an α value in the set $[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.2, 0.2, 0.5, 0.5, 0.4, 0.4, 0.6, 0.8, 1, 1, 1, 1]$ when a new node joins. The rest of the models will be the same. The resulting curve in Figure 4.5 shows that the saturation scaling parameter γ_s is almost the same with the result got by fixing α with the expectation of α, i.e, 0.3 in this case.

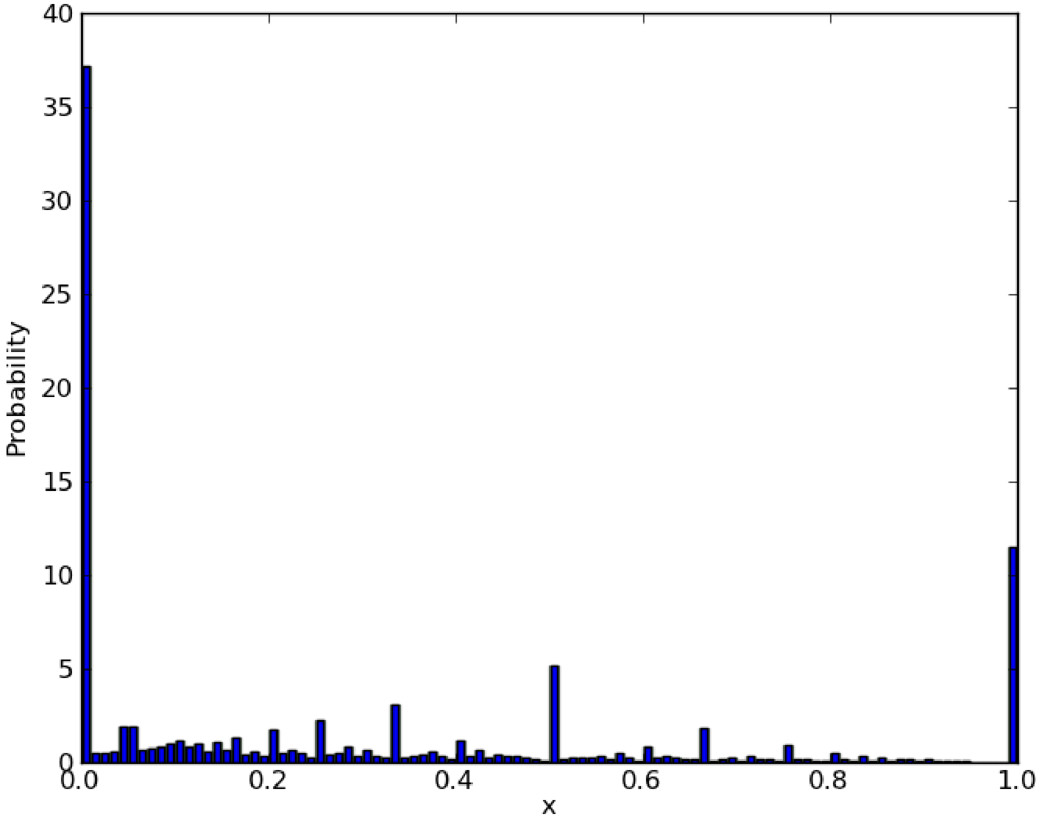


Figure 4.4: Distribution of α in empirical dataset $D1$

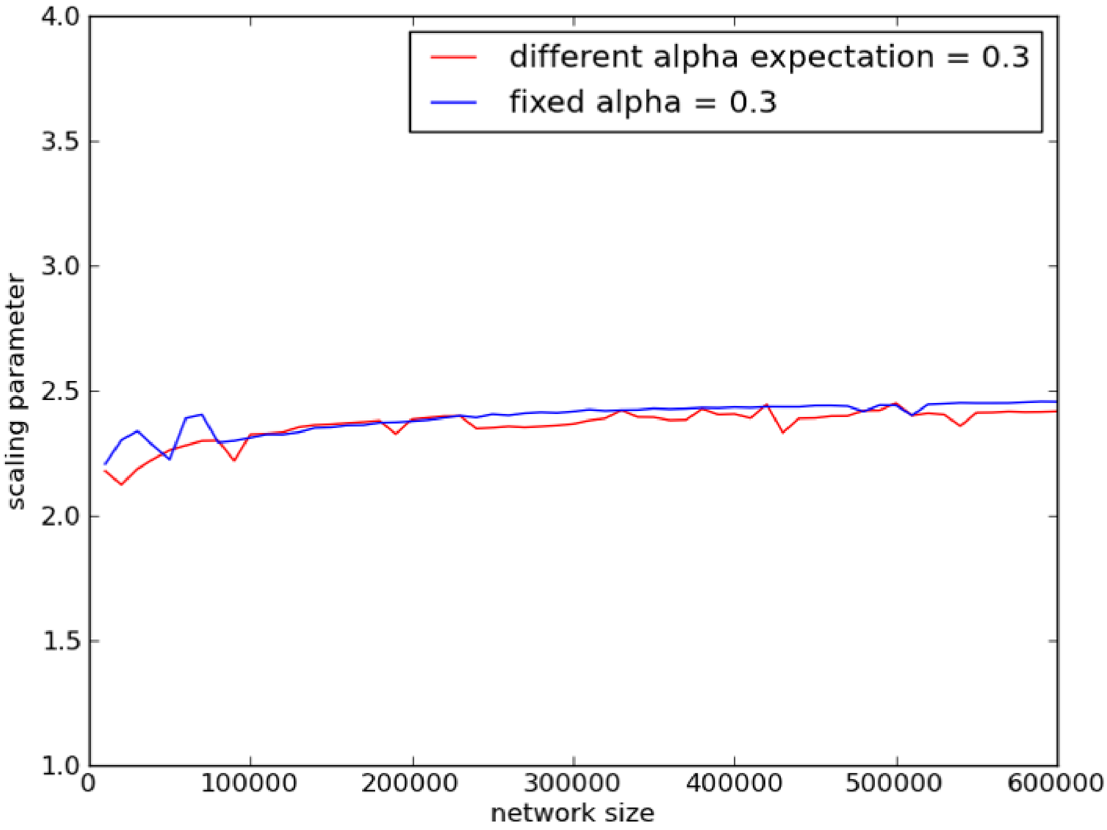


Figure 4.5: Fixed α compared with different α. Blue: α is fixed to 0.3 Red: α has a distribution similar to $D1$, with an expectation equal to 0.3.

4.3 Effect of Initial Network Size m_0

Figure 4.6 shows the effect of initial network size m_0 on the resulting scaling parameter during the network formation process. Basically the initial network size doesn't effect the saturation scaling parameter γ_s, but has an influence on the speed of reaching the stable stage. A large initial network size takes a longer evolution time to reach a saturation scaling parameter. This influence is quite obvious as shown in Figure 4.6 when m_0 equal to 80, the network didn't even reach a saturation stage before there are 0.6 million nodes.

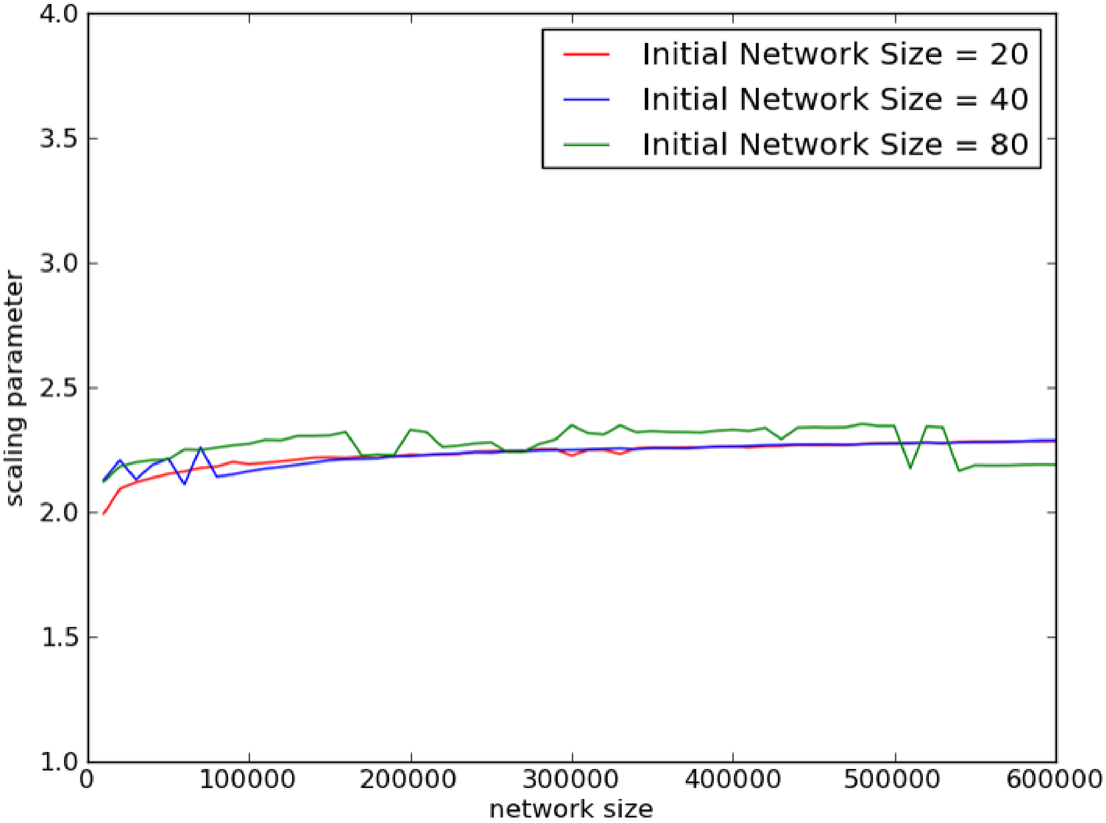


Figure 4.6: Effect of Initial Network Size m_0 on the result scaling parameter during evolution of the network

Chapter 5

Conclusions and Future Work

In this thesis the conclusion reached by previous researchers that Twitter network does not follow a power law distribution is challenged. A hypothesis that Twitter network contains two subnetworks following power law distribution is made. By extracting the social network and information network from two large scale empirical datasets and fitting them into power law distribution, the hypothesis is tested. Further, a two step configurable model that could generate a network with a similar structure as Twitter has been proposed. The validity of the model is then tested by mathematical analysis and simulations.

It is concluded in this thesis that the social ratio α is crucial in the formation process of such a network as Twitter, and a network with more social users has a larger value of α and a larger resulting scaling parameter γ for the information network. The structure of the social network part of Twitter influences the structure of the information network part, and to best describe its own formation process more human behavior related parameters should be taken into consideration.

This thesis provides a basic foundation for several lines of future works. With a better knowledge of Twitter network structure, the information propagation process and security issues can be analyzed quantitatively. On the other hand, the models proposed in this thesis could be further extended to represent general online social networks, which would be helpful in exploring the similarities and differences between different online social networks.

Bibliography

[1] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *arXiv preprint arXiv:1305.0215*, 2013.

[2] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM).*

[5] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[6] M. O. Jackson. *Social and economic networks.* Princeton University Press, 2010.

[7] M. O. Jackson and B. W. Rogers. Meeting strangers and friends of friends: How random are social networks? *The American economic review*, pages 890–915, 2007.

[8] H. Kwak, C. Lee, H. Park, and S. 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.

[9] MarketWatch. Twitters 10 best and worst moments. http://www.marketwatch.com/story/twitters-10-best-and-worst-moments-2013-11-05, Nov. 2013.

[10] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[11] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

[12] A. Sanzgiri, J. Joyce, and S. Upadhyaya. The early (tweet-ing) bird spreads the worm: An assessment of twitter for malware propagation. *Procedia Computer Science*, 10:705–712, 2012.

[13] B. Smith. Best times to post on social media. http://socialmediatoday.com/brianna5mith/1453951/best-times-post-social-media-infographic.

[14] Twitter. https://blog.twitter.com/2013/celebrating-twitter7, Mar. 2013.

[15] J. Weisenthal. This map shows the battle between the world's biggest social networks. http://www.businsider.com/map-global-social-media-competition-2013-12, Dec. 2013.

[16] Wikipedia. http://en.wikipedia.org/wiki/Power_law.

[17] O. Yagan, D. Qian, J. Zhang, and D. Cochran. Conjoining speeds up information diffusion in overlaying social-physical networks. *Selected Areas in Communications, IEEE Journal on*, 31(6):1038–1048, 2013.