Twitter Structure as a Composition of Two Distinct Networks

Meng Tong, Ameya Sanzgiri, Dimitrios Koutsonikolas and Shambhu Upadhyaya
Computer Science and Engineering, University at Buffalo, Buffalo, New York 14260
{mengtong, ams76, dimitrio, shambhu}@buffalo.edu

***Abstract*—Online social networks (OSN) such as Twitter and Facebook are playing important roles in our daily life, either to socialize and communicate with people or to share information. To better understand the information propagation in these networks, it is important to study their structure and formation process. However, to do so, it is imperative to analyze the roles these networks are used for – as news media or social networks. In this paper, we study the structure and formation of Twitter and attempt to answer what the role of Twitter is. We first examine the Twitter network creation process to gain insight into its structure. Our analysis indicates that the Twitter network can be formally modeled as a composition of two main networks that have different roles in information propagation. Following this we also propose a concise configurable two-step model that can generate a Twitter-like network to facilitate the development of a simulation platform for future research. We verify the validity of the proposed model by empirically analyzing two large datasets containing the topological information of Twitter and study its properties by means of mathematical analysis and simulation.**

I. Introduction

Online social networks (OSN), such as Facebook, Twitter, Instagram, etc., are becoming an important part of our daily lives. The popularity of these networks can be attributed to characteristics such as ubiquitous access and easy content production.

Twitter is especially interesting in this context since its content has been widely used in business and marketing [1], [2]. Further, the ability to provide real-time updates, as seen during "Arab Spring" and relief operations as seen after the Japan tsunami and earthquake, makes it an ideal platform for information dissemination. However, there have been cases where Twitter has been misused, such as the hijacking of the Associated Press account and subsequent bogus tweet, that resulted in the loss of millions of dollars in the stock market. To design mitigation techniques that prevent misuse of Twitter, it is imperative to understand the role of Twitter as well as the mechanisms of information propagation. However, to ensure that these techniques are realistic, scalable and practical, they need to be tested either by simulation or real experiments.

There is a lot of interest from the research community to understand the Twitter structure and information dissemination mechanisms [3]–[6]. However, the research is often dependent on datasets of Twitter structure which are not always easily available. This is further exacerbated by the lack of formal modeling or understanding of the Twitter structure, which is the main reason for the lack of suitable simulation platforms.

Two of the key challenges in creating a simulation platform are the quantification of the role of Twitter and the characterization of its structure. The public notion of Twitter is that it is an OSN used to interact with friends and also is a micro-blogging site to disseminate information [7]. The authors of [8] attempted to answer if Twitter is a social network or a news medium. Towards this end, they topologically analyzed a large dataset of the social graph of Twitter and one of their key findings was that the social graph did not fit Power Law distribution, a key attribute of social networks. However, they were not able to conclude with certainty the structure of Twitter.

The objective of this paper is to analyze the structure of Twitter based on the formation of its network and formally model its structure, in order to obtain a better understanding of the information propagation. Based on the purpose and formation of the Twitter network and its structure, we propose that Twitter is indeed a composition of two distinct networks and analyze them. We then propose a two-step configurable model that creates a Twitter-like structure while maintaining its properties and identify key parameters necessary for such a network. The model will serve as a first step towards creating a scalable simulation platform that can be used to analyze information propagation and user behavior.

The rest of the paper is organized as follows: The preliminaries and background appear in Sec. II. The key aspects of the proposed model are explained in Sec. III and empirically verified in Sec. IV. The simulation results are presented in Sec. V. Finally, we compare our work with related literature in Sec. VI before concluding the paper in Sec. VII.

II. Preliminaries

In this section, we first describe the way Twitter forms the network when a user creates a Twitter account. We then provide details on Power Law Distribution (PLD), including some properties as well as formation processes of networks that yield a PLD.

A. Twitter Network Formation Process

When a user creates a new Twitter account, a new network creation is initiated for the user, divided into three steps:

1) A list of popular users (e.g., celebrities or news media) is provided and the user is asked to "follow" five of them (need not necessarily be the suggested ones).

2) Twitter's Who to Follow algorithm [9] analyzes the areas of interest based on the selections in the first step. It then provides a categorized list so that the user can follow another five entities.

3) Twitter asks for permission to access the user's contact list (gmail or yahoo! mail) to find people that are already using Twitter and suggests five of them from the lists to follow.

The user → follower entity abstracts the dissemination model of information from a user to its followers, other users who "follow" a user. The information is propagated via Twitter specific messages called *tweets*. In terms of relationship, unlike other social networks, the relationship between a user and its follower in Twitter can be asymmetric. Specifically, when a user gains a follower, they both do not automatically follow each other, thus a user does not necessarily gain access to *all* the tweets of its followers.

B. Power Law Distribution

Power law distribution (PLD) is often used in the understanding and analysis of complex network structures. It holds a special significance in the purview of social networks.

Definition and Properties: Mathematically, the probability density function (pdf) of PLD can be expressed as

$$p(x) = Cx^{-\gamma} \tag{1}$$

where C is a constant and γ is the scaling parameter. It has been shown via real world datasets that the scaling parameter γ typically falls in the range between 2 and 3, although there are some exceptions [10].

Scaling Parameter: As can be seen from the equation, the scaling parameter governs the shape of the pdf curve. Basically, a smaller scaling parameter will lead to a more even distribution than a larger one. When the value of 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 to probabilities when value of the scaling parameters is equal to 2 or 3.

PLD and Social Networks: An important property is the heterogeneity among the possible values of a variable following a PLD. That is, a small number of values has high probabilities of appearance while the majority of values has very low probabilities of appearance. Thus, the average value cannot well describe the variables that follow a PLD. In the context of social networks, it can be used to analyze and gain understanding of the structure of the network. As has been shown in [11], many OSNs are likely to have a degree distribution following a PLD, with the scaling parameter falling in the range between 2 and 3, with the exception of Twitter [11]. Thus, one way to determine if a network is a social network, is to examine if the degree distribution follows PLD, a process known as distribution fitting.

Fitting of PLD: A simple way to judge whether a dataset follows a PLD is to plot its complementary cumulative distribution function (CCDF) on a log-log scale and examine if the plot is a straight line. Fitting PLD is a non trivial task [12] and while the method mentioned above is simple, it is not very accurate. This paper follows the method described in [12] which is an efficient implementation of the method in [10].

Preferential Attachment Model: This is one of the formation models that leads to a degree distribution that follows PLD. First described in [13], this model has two key features. First, the model assumes a growing network rather than forming links from a set of existing nodes; second, when a new node joins, it has a greater affinity to connect to popular nodes (higher degrees) than unpopular nodes, thus the name "preferential attachment." Let m be the number of links a new node forms upon joining and $d_i(t)$ be the degree of an already existing node i at time t, then when the new node joins, node i gains $m\frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)}$ new links.

To ease the modeling process, we use the mean field approximation approach by assuming that every new node forms the same number of links; this average behavior has been proved to be a good approximation [14]. This gives the increasing rate of $d_i(t)$ to be

$$\frac{dd_i(t)}{dt} = m\frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)} \tag{2}$$

Solving this differential equation with a start condition of $d_i(i) = m$ will give a PLD with a scaling parameter equal to 3. Though the original preferential attachment works on undirected networks and results in a PLD with scaling parameter equal to 3, a modification of the model can be used to incorporate directed networks and yields a wide range of values for the scaling parameter as shown in Sec. III-B1.

III. Analysis of the Twitter Structure

For the precise understanding of the structure as well as the role of Twitter, the authors of [8] analyzed the topology of Twitter users and concluded that Twitter exhibited a non-power-law follower distribution, a short effective diameter, a low reciprocity, which all mark a deviation from known characteristics of human social networks. Quantifying the role of Twitter is also a difficult question since it is largely user-dependent, which means different users use Twitter differently. However, the formation process as explained in Sec. II provides some insight into analyzing the network.

The process almost[1] clearly separates the formation of two subnetworks, say, information network and social network. The first two steps in Sec. II-A can be regarded as helping the new user form the "information network", by suggesting popular users. This is very reasonable since users want someone with public trust/credibility as information sources, thus popular users provide good choices as they are trusted by a large number of users. The formation of the information network is also a basic characteristic of the preferential attachment model which is further displayed by the "find and follow well-known people" as part of the second step.

The third step builds the "social network", by importing from other existing social relationships, typically people from the email contact list who are already using Twitter. The formation process also reveals two other important parameters that can be helpful in building a model of the Twitter structure. The first one is the total number of users that a user will follow upon joining Twitter, which if the user strictly follows the Twitter suggestions, is 15. The second one is the ratio of the number of users a new user follows by searching his contact list to the total number of users the new user follows, which we denote as α. This parameter α is defined as the "social ratio" in this paper and the significance of this will be explained in Sec. V. Thus, if a new user strictly follows the Twitter suggestions, α will be $1/3$, as 5 out of 15 of Twitter users followed by the new users are supposed to be "real friends."

[1] By almost we mean that there is a possibility for some of the users' contacts to appear in the suggested lists of steps 1 and 2 of Sec. II-A.

In the following section we use the insights from the formation process to analyze the Twitter network.

A. Network Separation

Based on the preceding discussion, we hypothesize that the Twitter network is separable into two different networks based on the usage purposes. Further, since the original Twitter follower network could be regarded as a mixed-purpose network or a combination of a social and an information network, it is reasonable that it does not strictly follow a PLD as concluded by the authors of [8].

In view of this, the two subnetworks extracted from the Twitter network can be formally defined as follows:

- **Social Network**: a network containing all mutual relationships. This is an undirected graph where every pair of connections implies that the connected users mutually follow each other on Twitter. Nodes only in the social network correspond to the white circular node in the social network in Fig. 1. These nodes have only mutual relationships.
- **Information Network**: a network containing all the one-way relationships. This is a directed graph where every pair of connections implies that one user follows the other but not vice versa. Nodes in the information network correspond to the triangular nodes in the information network in Fig. 1 and have only a one-way relationship.

There are also those nodes that exist in both networks and have both mutual as well as one-way relationships. These correspond to the shaded nodes in Fig. 1. It should be noted that if two nodes are connected in the social network and both appear in the information network, they will not be connected in the information network.

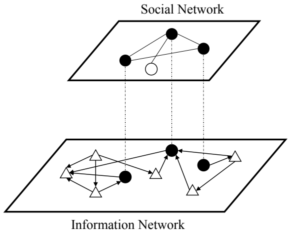


Fig. 1. Network Separation. Top: Social Network. Bottom: Information Network. Black round nodes appear in both networks, whereas triangular nodes appear only in the information network and white round nodes only in the social network. An arrow from node A to node B in the information network indicates user A follows user B.

The three types of nodes correspond to the three different types of users discussed in the previous section. The nodes in these two networks could be overlapping but the links in these two networks are mutually exclusive. Thus, a link cannot appear in both networks. It should be noted that only the follower network is considered here, i.e., the out degrees of all the nodes are considered. The reason is that a user's tweets will appear in all its followers' timeline, but has no effect on its friends' (users it follows) timelines. From the perspective of information propagation, the out degree of a node typically indicates how many other nodes it could reach in one hop, i.e., its ability to spread information from the point of view of size. Hence, 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 the set that contains all the nodes appearing in the Twitter follower network, $V_a = \{v|d(v) \geq 0\}$, where $d(v)$ is degree (in and out) of node v. As users with no followers have no ability to spread information and cannot have negative followers, it is obvious that all $v \in V_a$ have a non-negative number of followers. E_a is the set containing all the follower relationships in the network. If $e_{ij} \in E_a$, then user i follows user j.

We can then define the social network as the graph $G_s = (V_s, E_s)$ and the information network as the graph $G_i = (V_i, E_i)$, with the following relationships:

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

Based on the idea of network separation, it is meaningful to investigate if the two subnetworks have a more clear degree distribution. Fundamentally, we want to verify if either one or both of them would be a better fit of the PLD, thus conforming to the notion of a human social network. The testing of our hypothesis is described in Sec. IV.

B. Generation of Proposed Models

Based on the observations from the real world process in Sec. II-A and the above analysis, we propose two configurable two-step formation models to generate a network capturing the degree distribution of the real Twitter network. It is important to note that the goal is not simply to generate a network with its degree following PLD; but rather to find a process which is similar to Twitter user behavior as much as possible and could lead to a similar network distribution at the same time while making it scalable.

1) Description of Proposed Models: At each time step, a new node joins a network, making the model a growing network formation. Upon joining the network, the new node selects a subset of nodes, m, from the existing nodes with whom to form a relationship. Among these m nodes, some are selected as information sources, while others are selected as friends in the real world. The former ones (step 1) will appear in the information network and are named "information network nodes." The new node forms a directed link with each of these nodes; such links are called "information links." The latter ones (step 2) will appear in the social network and are named "social network nodes" or "mutual followers." These nodes will form two directed links, from the new node to each of them as well as in the opposite direction; such links are called "social links." We assume that all the following relationships are formed when a new node joins the network, which, while not being realistic, is used to simplify the model as considering forming links between existing nodes would be equivalent to changing some initial parameters in our proposed model. Based on this assumption, there are $(1-\alpha)m$ information network nodes and αm social network nodes, where α is the social ratio.

The following two steps are used by a node to connect to the m nodes.

Select information network nodes. Based on the preferential attachment scheme, the probability of an existing node to be selected as an information network node by the new user is directly proportional to its current in-degree. This process is similar to Twitter suggesting new account users to follow popular users. However, it should be noted that the current in-degree of an existing node includes both its information network in-degree as well as its social network in-degree. This can be reasoned from two perspectives. From the existing node's point of view, the social network in-degree also contributes to its popularity, e.g., viral videos, photographs, etc. Similarly, from the new node's point of view, when deciding whether to follow another user as an information source, it is based completely on the user's out degree, which is an aggregate of both mutual followers and pure followers.

Select social network nodes. 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. The modeling of this can be done via two processes.

(i) *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 and the degree distribution of this model should follow a PLD.

(ii) *Multiplicative process*: In [15] 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 and is referred to as Model II.

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

In the real world scenario, the values of m and α vary for all users. However, to simplify the modeling, the mean field approximation is used in these models assuming that, every user behaves like an average user with the same number of nodes to connect to, and with the same friends to select. It will be shown later via simulation that the behavior of α may only be effected by its expectation. These models are configurable in the sense that α can be adjusted. We now sketch a mathematical analysis for the above two models.

2) Mathematical Analysis: We assume that $d_k^i(t)$ is the in-degree of node k in the information network at time t; similarly, $d_k^s(t)$ is the in-degree of node k in the social network at time t, and $d_k(t)$ is 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

$$\frac{dd_k^i(t)}{dt} = (1-\alpha)m\frac{d_k^i(t)+d_k^s(t)}{\sum_{j=i}^{t} d_j(t)} \tag{3}$$

Similarly, the number of new social links an existing node k will gain, assuming Model I is:

$$\frac{dd_k^s(t)}{dt} = \alpha m\frac{d_k^s(t)}{\sum_{j=i} d_j^s(t)} \tag{4}$$

Solving this equation for the social network,

$$d_k^s(t) = \alpha m(\frac{t}{t_k})^{0.5} \tag{5}$$

where t_k is the time at which node k was added to the system.

Substituting (5) back to (3) gives the rate of increase as

$$\frac{dd_k^i(t)}{dt} = \frac{d_k^i(t)}{At} + \frac{\alpha m}{At_k^{0.5}} \times \frac{1}{t^{0.5}} \tag{6}$$

where $A = \frac{1+\alpha}{1-\alpha}$. Solving the differential equation for the information network gives

$$d_k^i(t) = \frac{2\alpha m}{2-A} \times ((\frac{t}{t_k})^{\frac{1}{A}} - (\frac{t}{t_k})^{0.5}) \tag{7}$$

Comparing with (5), we can observe that there are two power law components (consistent with our hypothesis) and that the resulting scaling parameter is affected by α as well as the network structure of the social network part. This means that a larger α will lead to a larger scaling parameter of the information network. Similarly, for Model II the change in number of social links will be given by:

$$\frac{dd_k^s(t)}{dt} = \frac{\alpha m}{t} \tag{8}$$

The above analysis is similar to the one in [13] and is customized to our context. We omit the details of the derivations for the sake of brevity.

IV. Empirical Study

In this section, we use two datasets to empirically validate our proposed model.

A. Dataset Information

The Twitter datasets from [8] and [16] are used in this paper, and are denoted as $D1$ and $D2$, respectively. Table I lists the basic information from the two datasets.

From Table I, we can see that the number of Twitter users increased by more than 25% in just three months. Further, it should be noted that these two datasets contain the whole topology of Twitter network at the time they were crawled. The datasets provide a measure of ground-truth of Twitter network, since a large network like Twitter is shaped by all its users. Thus any conclusions reached from a partial or sampled dataset would not be convincing enough. Further, by studying the difference of the two datasets crawled in a close time period, the evolution trend of Twitter network could be more clear, thus providing insights into the topological analysis.

B. Fitting and Results

To test our hypothesis using real data traces, the social network and the information network are first extracted from the originally unseparated datasets. To our surprise in both the datasets there are only about 50% of Twitter users who have at least one mutual link with another user. This corroborates the conclusion reached in [8] that Twitter has a lower level of reciprocity. The value of reciprocity of 22% in $D1$ (21.6% in $D2$) is low when compared to other OSNs like Flicker and Yahoo! 360. The difference between the maximum degree and the average degree in the social network also suggests that the heterogeneity among nodes could possibly lead to the social network fitting a PLD. Table I presents the statistics for the information and the social networks. From the table we can see that in the information network only a small fraction of Twitter users have no followers and that the difference between the average degree and maximum degree is significant. Although $D2$ is larger in size than $D1$, both exhibit the same basic properties in the separated networks, indicating that Twitter may have stepped into a stable stage in its evolution.

TABLE I. Twitter Network Statistics

Attribute	$D1$	$D2$
Total users	41,652,230	52,579,682
Total Links	1,468,365,182	1,963,263,821
Social Network Characteristics		
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
Information Network Characteristics		
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

The next step is to check whether the degree distribution of these two subnetworks follows a PLD. We first plot the CCDF in normal scale and log-log scale. As mentioned in Sec. II-B, the 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 2 and 3 show the plot of the CCDF of $D1$; the CCDF plots of $D2$ are similar and hence are not shown.

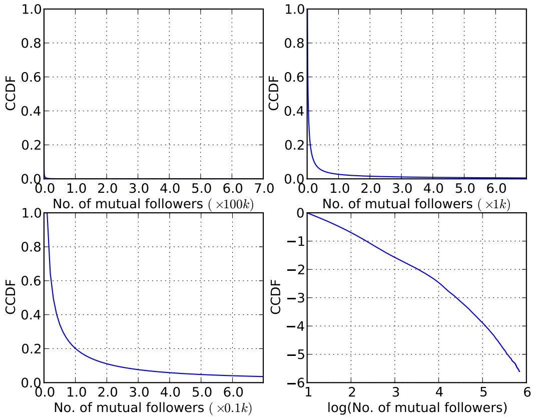


Fig. 2. CCDF of Social Network Degree in $D1$ Top Left: CCDF in normal scale. Top Right and Bottom Left: Zoomed versions (100x and 1000x). Bottom Right: CCDF in log-log scale

In Fig. 2 and 3, the log-log plots are very close to a straight line, starting from a lower bound, thus indicating that our hypothesis has a high probability to be true. In order to confirm the hypothesis, the exact scaling parameters need to be calculated. Following the process described in [12], the datasets are fitted into PLD, and the scaling parameter as well as the goodness of fit compared to other candidate heavy-tail distributions (exponential/lognormal), are calculated.

The fitting results are shown in Table II, where the comparison with two alternative distributions is shown by the unnormalized likelihood ratio of PLD to exponential and lognormal by columns three and four respectively. As described in [12], a positive value indicates that the fit follows PLD, whereas a negative value indicates that the fit follows the other heavy-tail distribution, namely, exponential or lognormal.

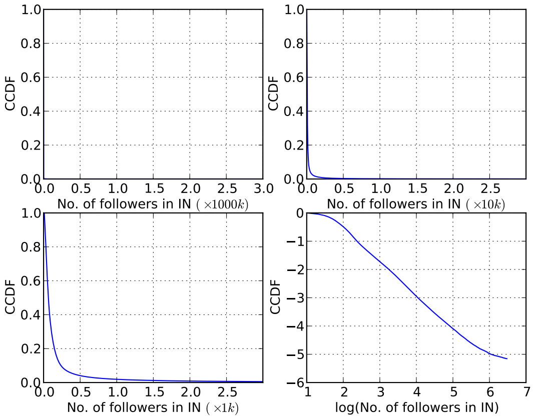


Fig. 3. CCDF of Information Network Degree in $D1$ Top Left: CCDF in normal scale. Top Right and Bottom Left: Zoomed versions (100x and 1000x). Bottom Right: CCDF in log-log scale

TABLE II. Power Law Fitting of Social & Information Network

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

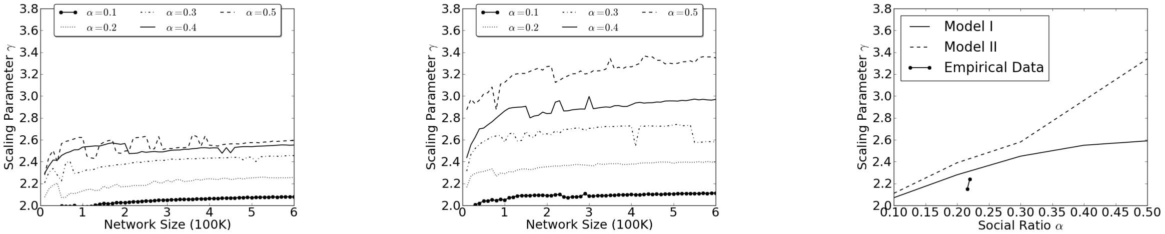
Basically, the information network is a good fit of PLD with the scaling parameter equal to 2.24 in $D1$ and 2.15 in $D2$ compared to the exponential and lognormal distributions. However, for the social network the power law fitting does not showcase a better fit over lognormal distribution. In fact, as the fitting algorithm is not deterministic, the power law is a better fit to the lognormal only some of the times. This is the reason why two different models are tried for these two possibilities in Sec. III-B1.

V. Simulation and Results

In this section, we conduct simulations to analyse the various parameters that affect the formal modeling of Twitter structure.

Simulation Setup: All simulations start with an initial network containing m_0 nodes, fully connected with each other, in order to mimic the launching process of Twitter and other online products. Basically, they would start with invitation or internal test, which indicates highly connected relationships between the initial users. As described in Sec. III-B1, the formation of the network continues by adding one node at each time step. Upon joining the network, the new node selects information type and social type nodes to connect to, and the network gets updated. If a node is selected as one of the two types, then it cannot be selected again. All simulations stop when the network size reaches 0.6 million, which is an empirical value as after this the network structure is observed to be stabilized.

Effect of Fixed α: Fig. 4(a) and Fig. 4(b) show the effect of α along the way of the network evolution, for the two models proposed in Sec. III-B1. In these simulations, m and m_0 are both set to be 20. Different values of α are tested on both



(a) Effect of α on γ (Model I)

(b) Effect of α on γ (Model II)

(c) γ_s for the models and the two datasets

Fig. 4. Effect of social ratio on scaling parameter.

models and the values are selected so that αm and $(1-\alpha)m$ are both integers.

Although there are some oscillations in the curves, in general, all of them show the same trend for the 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. Fig. 4(c) shows the saturation scaling parameter γ_s for different models as well as for the empirical datasets $D1$ and $D2$ [2]. It can also be observed from the figures that a larger α produces a larger scaling parameter γ, which is consistent with the mathematical deduction.

To compare the two different models, closer attention should be paid in Fig. 4(a) and Fig. 4(b) to the lines when the "social ratio" α is equal to 0.2 since they are closest to the overall α in the empirical dataset, which is reported to be 0.22 in [8]. In Fig. 4(c), 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 $\alpha = 0.22$ (in $D1$) and 2.15 when $\alpha = 0.216$ (in $D2$), it can be concluded that Model I is a good enough fit for generating the desired information network.

However, the fitting of the social network part remains an open question. As analyzed in Sec III-B2, the social network structure has an impact on the fitting scaling parameter of the information network while by 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 (from Fig. 4(c)), it is reasonable to make the hypothesis that a social network with a lower scaling parameter will further decrease the saturation scaling parameter γ_s progressively, yielding a value even closer to 2.24 or 2.15. 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 considered.

It is interesting to think about the role that the "social ratio" α plays in the network formation process and the resulting structure. Generally, it reflects on average how "social" the users in the network are, i.e., how often or how willing the users are to use it as a social network with their friends. Thus, α can be regarded as an intrinsic property of a particular network, determined by the nature of the network. From the two empirical datasets, we calculated that the social ratio α is decreasing, from 0.22 in June 2009 to 0.216 in September 2009; this indicates that users are getting more information-driven on Twitter and use it more as an information source. 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 discussion above, we conclude that Model I is the better of the two models and all further simulations were performed with Model I.

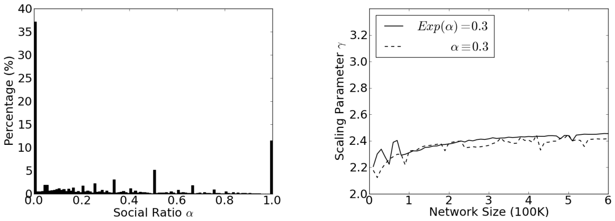
Different α for Different Users: In Sec. II we assumed in our models that all users have the same social ratio α, under the mean field approximation approach. We test the validity of our assumption as well as the goodness of this approximation, by assigning to α values from a distribution similar to that calculated in the empirical dataset, as shown in Fig. 5(a). It can be observed that a large fraction of the users have no or very small percentage of mutual friends, and some values of α are more frequent than others. This distribution is simulated when the new node joins, 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]. The resulting curve in Fig. 5(b) shows that the plot of the saturation scaling parameter γ_s is almost the same when all the nodes have the α value fixed at 0.3, and when all the nodes have a value of α drawn from a distribution with an expected value of $\alpha = 0.3$.

Effect of Initial Network Size: Fig. 6 shows the effect of initial network size m_0 on the resulting scaling parameter during the network formation process. We observe that the initial network size does not affect 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 time to reach a saturation scaling parameter. This influence is quite obvious in Figure 6; when m_0 is equal to 80, the network did not even reach a saturation stage before there are 0.6 million nodes.

VI. Comparison to Related Work

The analysis based on topological datasets has led to inference on followers, friends, geographic distributions, etc. [3]–[5]. Similarly, the analysis based on the content-based datasets has revealed aspects of information propagation such as the type of content retweeted, the users that retweet information, etc. [1], [6], [17]. These works also try to predict the type of content based on the real-world dataset. The work in [18] uses both types of datasets to understand the structure and interaction between students who have used Twitter but it does not provide a formal model. The authors of [19] use sociology concepts to understand the network structure and analyze how it influences users to "un-follow" other users. Our work differs

[2] The two points correspond to the γ_s values calculated by obtaining the α values from datasets $D1$ and $D2$.



(a) Distribution of α in $D1$

(b) Fixed α compared with different α

Fig. 5. Analysis of social ratio.

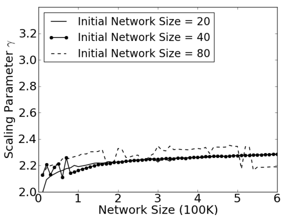


Fig. 6. Effect of Initial Network Size m_0 on γ as the network evolves

from the rest in two aspects. First, in terms of methodology, we use a combination of analyzing the formation process to propose a model and empirical validation. Second, we present a concise two-step model that can generate a structure similar to the Twitter network. Finally, we also identify the parameters that can affect the structure of the network.

VII. Conclusions and Future Work

In this paper, the conclusion reached by previous researchers that the Twitter network does not follow a PLD is challenged through mathematical and empirical analysis. A hypothesis that the Twitter network contains two subnetworks following PLD is made and a formal model of the Twitter structure is presented. Further, a two-step configurable model that could generate a network with a similar structure as Twitter has been proposed. The hypothesis is validated by extracting the social network and information network from two large scale datasets and fitting them into PLD. Finally, we also identify some parameters and test their effect on the Twitter structure via simulations. We show 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.

The results of this paper provide a basic foundation for several lines of future research. With this formal model of the Twitter network structure, information propagation process and security issues can be analyzed quantitatively. Conversely, the models proposed in this paper could be further extended to represent general OSNs, which would be helpful in exploring the similarities and differences between different OSNs. Our goal with the analysis and modeling presented in this paper is to facilitate the building of a simulation platform for OSNs that will help the research community in their research with such networks.

Acknowledgment

This research is supported in part by National Science Foundation Grant No. DUE-1241709. The authors also acknowledge Abhishek Cumbakonam Desikan for his help with this manuscript.

References

[1] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *SocialCom, IEEE*, Aug 2010, pp. 177–184.

[2] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the American society for information science and technology*, vol. 60, no. 11, pp. 2169–2188, 2009.

[3] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: Understanding microblogging usage and communities," in *Proc. of the 9th WebKDD and 1st SNA-KDD*, 2007, pp. 56–65.

[4] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," *arXiv preprint arXiv:0812.1045*, 2008.

[5] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about twitter," in *Proceedings of the first workshop on Online social networks*, 2008, pp. 19–24.

[6] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Proc. of WWW*, 2011, pp. 113–114.

[7] Wikipedia, "Twitter," http://en.wikipedia.org/wiki/Twitter, April 2014.

[8] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proc. of 19th WWW*, 2010, pp. 591–600.

[9] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, "Wtf: The who to follow service at twitter," in *Proc. of the 22nd WWW*, 2013, pp. 505–514.

[10] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[11] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. of the 7th ACM SIGCOMM IMC*, 2007, pp. 29–42.

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

[13] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

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

[15] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet mathematics*, vol. 1, no. 2, pp. 226–251, 2004.

[16] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proc. of ICWSM*, Washington DC, USA, May 2010.

[17] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern, "Predicting information spreading in twitter," in *Workshop on computational social science and the wisdom of crowds, nips*, vol. 104, no. 45, 2010, pp. 17 599–601.

[18] K. Stepanyan, K. Borau, and C. Ullrich, "A social network analysis perspective on student interaction within the twitter microblogging environment," in *ICALT IEEE*, 2010, pp. 70–72.

[19] F. Kivran-Swaine, P. Govindan, and M. Naaman, "The impact of network structure on breaking ties in online social networks: Unfollowing on twitter," in *Proc of SIGCHI*, 2011, pp. 1101–1104.