The Effects of Restrictions on Number of Connections in OSNs: A Case-Study on Twitter

Saptarshi Ghosh
CSE, IIT Kharagpur, India
saptarshi@cse.iitkgp.ernet.in

Gautam Korlam
CSE, IIT Kharagpur, India
gautam@cse.iitkgp.ernet.in

Niloy Ganguly
CSE, IIT Kharagpur, India
niloy@cse.iitkgp.ernet.in

Abstract

Most popular Online Social Networks (OSNs) in today's world, such as Facebook, Orkut and Twitter impose restrictions on the number of friends / connections that a member can have in the network. This is primarily due to two reasons - to limit spam and to reduce the strain on the system due to member-to-all-friends communication. We study the effects of such restrictions on node-degree, on the topological properties of the OSN networks, taking the restriction imposed by Twitter as a case-study. To the best of our knowledge, this is the first study of its nature, on any OSN. Towards this end, we use a network growth model based on preferential attachment to develop an analytical framework that can be used to assess the effects of various forms of restrictions on OSNs, as well as to design new restrictions of varying rigidity.

1 Introduction

Online Social Networks (OSNs) are among the most popular sites on the Web in the present times, and the well-known OSNs, such as Facebook, Twitter, Orkut, Flickr and so on, each have over 50 million users (or members, the terms will be used interchangeably) presently. With the rapidly increasing member count, the successful OSNs have been facing a number of challenges over the past couple of years [20]; one of them is spam and other malicious activities by certain members.

Spammers typically use the member-search features provided by the OSN to contact (establish friendship links with) a large number of members and then use the methods of communication provided to send spam, thus annoying the legitimate users of the OSN. If not controlled, the amount of spam may rise to a level that prompts large numbers of legitimate users to leave the OSN.

Several popular OSNs have adopted a common technique to counter spam and improve the experience of legitimate users: they have imposed restrictions on the number of friends / connections that a user can have in the network. For example, the number of friends that a user can have is restricted to 1000 in Orkut and 5000 in Facebook. Flickr restricts the number of non-reciprocal contacts of members to 3000. Twitter has placed a more intelligent limit [4] on the number of people that a member may 'follow', as explained in section 2. These limit-based restrictions act as a first line of defence in controlling spam.

Apart from preventing the spammers from contacting other members indiscriminately, such limit-based restrictions also serve another important purpose. Popular OSNs have been suffering from scaling issues, due to the steady increase in their membership, which causes these sites to often have high latency [20]. Most of these OSNs, like Facebook and Twitter, offer features for real-time one-to-many communication, i.e. a user is allowed to post messages that would be communicated to all friends of that user in real-time. Hence, if users are allowed to have an excessive number of friends, the large number of message communications required may adversely affect the performance of the system.

However, the restrictions on the number of friends, as imposed by several OSNs, affect not only the spammers but also the legitimate users of the networking service. Thus such restrictions are criticised by a fraction of the legitimate users of the OSNs, as an encroachment on the freedom of users to have more friends [9]. Moreover, a systematic understanding of the relation between the restrictions and the desired performance improvement is also missing.

Hence, for the design of effective limits, an analysis of the dynamics of the creation of links in the social network (by legitimate users and spammers) and an understanding of the effects of different forms of restrictions on these dynamics is required. An analytical framework modeling the node / link creation and the restrictions, and predicting the emergent degree distributions can be an efficient method to gain this understanding; this is what

this paper does. This paper proposes a general framework to model the effects of restrictions on node-degree, on the topological properties of OSNs. The restriction imposed by Twitter [1] is taken as a case-study where empirical data collected by crawling the Twitter network show the effect of this restriction as a spike and a decay in the out-degree distribution.

Several studies have been conducted on the growth dynamics in OSNs and the topological characteristics of the network that emerge as a result of these dynamics [15, 17]. However, the changes in the topological characteristics of OSNs, due to imposed restrictions on node-degree, have not been formally studied or modeled till date, to the best of our knowledge. The effect of 'hard' cut-offs on node-degree have been studied in peer-to-peer networks where the number of connections that a peer can accept is limited by the finite bandwidth of the peer node [11, 18]. Unlike peer-to-peer networks, the Twitter social network is directed, and the imposed cut-off is only on the out-degree of nodes. Moreover, this cut-off is a 'soft' one and can be overcome by nodes which have 'sufficient' in-degree. Hence a completely different set of mathematical tools need to be developed to explain the emerging degree distribution from such dynamics.

The rest of the paper is organized as follows. The restriction imposed by Twitter is detailed in section 2. Section 3 describes the procedure used for crawling the Twitter network and the characteristics of the empirical data collected are discussed in section 4. A network growth model based on preferential attachment [8] is modified by incorporating a restriction similar to that in Twitter in section 5 while the observations drawn using the model are given in section 6. Discussions and conclusions of the study are drawn in section 7.

2 The Twitter Follow-Limit

Twitter [1] allows members to communicate among each other through the exchange of short messages called 'tweets' (each of 140 characters or less) and to form a social network, based on interest of a member in the tweets of another. If a Twitter user u finds another user v's profile or tweets interesting, u can "follow" v, by which, tweets posted by v will be made available to u. In Twitter terminology, if user u follows user v, v is said to be a "friend" of u and u is said to be a "follower" of v. The following relationship on Twitter is not mutual i.e. u follows v does not necessarily imply v follows u.

In graph-theoretic terms, the Twitter social network is a directed network where members are represented as nodes, and nodes u and v are connected by a directed edge $u \rightarrow v$ if member u follows member v. In this model, the number of friends of a member u is analogous to the out-degree of the node u, and the number of followers of u is analogous to the in-degree of u.

Thus, in the Twitter network, the out-degree of u (i.e. the number of members whom u follows) can be thought of as a measure of u's social activity or her interest to collect information from other members. Similarly, the in-degree of u is a measure of the popularity of u in the Twitter social network: this is the number of other members who are interested in the tweets posted by u.

According to analysis [19] carried out on Twitter in October 2009, Twitter has experienced an exponential growth in membership starting from the later part of the year 2008, making it one of the most popular OSNs today. This growing popularity of Twitter has attracted the attention of spammers who attempt to manipulate the features provided by Twitter to gain some advantage, such as driving Twitter users to other websites that they (i.e. the spammers) post as links in their tweets. One technique commonly adopted by spammers to gain attention is to indiscriminately follow numerous other users, in the hope of getting followed back; this technique is termed as "Aggressive Following" or "Follow Spam" [5].

To limit follow spam and to reduce the strain on the website [4], Twitter enforced a restriction on the number of users that a user can follow (i.e. on the out-degree), in August 2008 [5]. Every user is allowed to follow up to 2000 others, but "once you've followed 2000 users, there are limits to the number of additional users you can follow: this limit is different for every user and is based on your ratio of followers to following.", as given in the Twitter Support webpages [4].

However, Twitter does not specify the restriction fully in public [5]. In the absence of official specification, there have been several conjectures regarding the Twitter follow-limit [6, 7]. We here mention two commonly accepted versions. Let the number of followers (in-degree) of member u be denoted by u_{in}. Then the maximum number of members whom u can herself follow (maximum possible out-degree), denoted by u_{out}^{max}, is:

- version 1 (known as the '10% rule'):

$$u_{out}^{max} = \max\{2000,\ 1.1 \cdot u_{in}\}$$

- version 2:

$$u_{out}^{max} = \begin{cases} 2000 + 0.1 \cdot u_{in} & \text{if } u_{in} < 2000 \\ 1.1 \cdot u_{in} & \text{if } u_{in} \geq 2000 \end{cases}$$

Both versions imply that if a user wants to follow (out-degree) more than 2000 people, she needs to have at least a certain number of followers (in-degree) herself. A closer look shows that version 1 is a much stringent restriction as compared to version 2. For example, if u is already following 2000 members and wants to follow

one more, u requires to have at least 1820 followers by version 1, but just 10 followers by version 2. If u herself has more than 2000 followers, both versions behave identically by limiting the number of people that u can follow to 110% of the number of followers of u.

3 Methodology for Data Collection

We collected empirical data of the Twitter network using the Twitter API functions [2]. The Twitter social network has now grown to an extent (more than 50 million users) that makes collecting the entire network practically infeasible (contrary to what could be done [12] in 2007). Moreover, the collection of Twitter network data is also constrained by the rate limits enforced by Twitter: at most 150 API calls can be made in an hour [3]. Hence recent studies on Twitter [19] have to resort to obtaining only a sample of the Twitter network.

We used the Twitter API to collect the information of users by a breadth-first search (BFS) starting from a designated user in the network (also known as the snowball sampling method). The duration of data gathering was from October 23 to November 8, 2009. The BFS was continued until 1 million unique users were discovered; this seemed to us to be a reasonable sample size to be a representative of the entire network. The profile information collected for each user includes her number of friends, number of followers, number of tweets posted and other information such as the date of creation of the account and her geographical location.

It has been demonstrated in [16] that a property of the sample obtained by the partial BFS crawl method employed by us can be estimated to be similar to the corresponding property of the entire network, if the property of interest reaches a stable regime as the size of the sample grows during the measurement. We have verified that the properties of interest in this paper, i.e. the in-degree and out-degree distributions, of the first 25%, 50% and 75% of the network crawled by the BFS sampling technique are very similar to those of the total sample. Hence it may be concluded that the properties of the degree distributions of the crawled sample, as discussed in the next section, are likely to be similar to those of the entire Twitter network.

We also conducted a set of separate, smaller crawls of the Twitter network in the months of January and February, 2010. These crawls were started from randomly selected nodes (i.e. Twitter users), and each crawl was configured to crawl up to 50,000 nodes. We found that the in-degree and out-degree distributions of each of these smaller samples preserve the major characteristics of those of the largest sample.

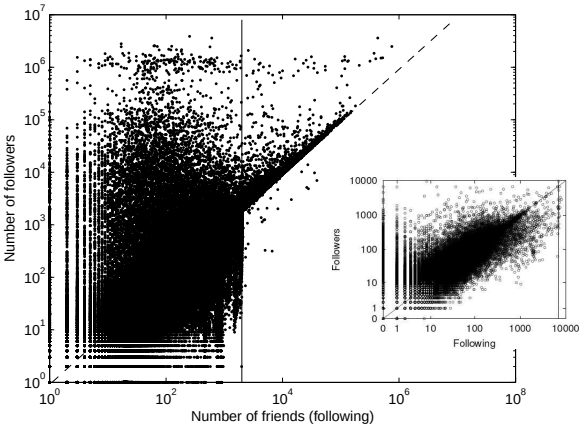

Figure 1: Scatter plot of number of followers and number of friends of Twitter users (a) Data collected in Oct-Nov, 2009, along with the lines $x = 1.1 \cdot y$ and $x = 2000$ (b) (inset) Data collected in Jan-Feb, 2008 (reproduced from [14])

4 Characterization of Twitter

This section discusses the statistics of the number of followers (in-degree) and friends (out-degree) of users in the Twitter social network, as obtained from the crawled empirical data. The effect of the restriction imposed by Twitter is clearly demonstrated through these statistics.

Scatter plot

Fig. 1 shows the scatter plot of the followers / friends spread in the Twitter dataset obtained by our crawl in October-November 2009. To exhibit the effect of the imposed restriction, the scatter plot obtained from the data collected in January-February 2008 (which was before the restriction was enforced) is reproduced from [14] in fig. 1(b) (inset).

Several changes in the character of the Twitter social network can be clearly identified from the scatter plots in fig. 1. First, an idea of the recent exponential growth in the size of the Twitter network can be obtained by comparing the maximum values on either axis in fig. 1(a) and fig. 1(b) (inset): whereas the maximum follower-count and friend-count was close to ten thousand in early 2008, the corresponding values are over 1 million in late 2009.

Secondly, the scatter plot in 2008 (fig. 1 inset) is seen to be symmetrical about $x = y$ for the entire range of x (number of friends), but the scatter plot in 2009 has a sharp edge at the abscissa corresponding to 2000 friends. This is a consequence of the restriction imposed by Twitter on the number of friends - only a small fraction of members (about 6.68% in our dataset) have more than 2000 friends. It is also evident from fig. 1 that the members who have more than 2000 friends need to have a sufficient number of followers, such that their number

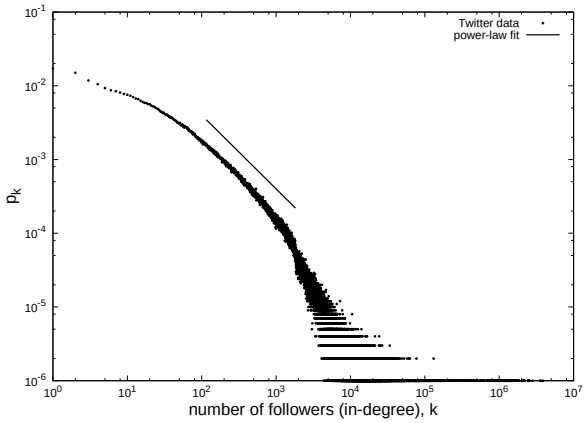

Figure 2: Distribution of number of followers (in-degree) in Twitter and power-law fit $p_k \sim k^{-1.0}$ below 2000

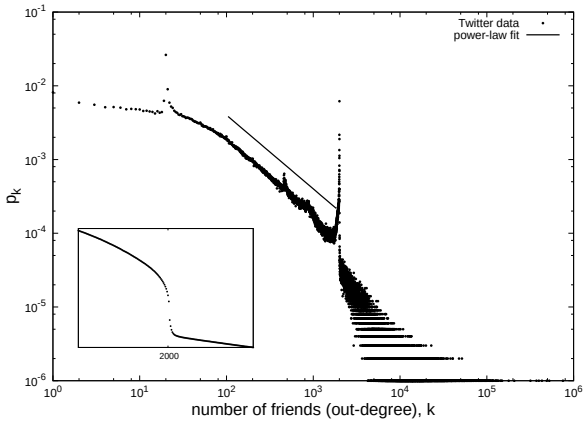

Figure 3: (a) Distribution of number of friends (out-degree) in Twitter and power-law fit $p_k \sim k^{-1.0}$ below 2000 (b) (inset) The discontinuity in the cumulative out-degree distribution around 2000

of friends remains less than 110% of their followers; the data points corresponding to most of these users lie to the left of the $x = 1.1 \cdot y$ line, verifying the '10-percent rule' explained earlier.

It is seen from fig. 1(a) that there exists a small fraction of members (less than 0.4%) in the crawled dataset who seem to violate the restriction, specially at higher values of y (= number of followers): these members correspond to the data points lying to the right hand side of the $x = 1.1 \cdot y$ line. On verifying these accounts in our dataset, it is seen that several of these member accounts have been created before the restriction was imposed. For the others, it seems that Twitter allows some relaxation of the restriction for popular members who have a relatively large number of followers, on case by case basis. The lack of official specification of the restriction from Twitter disables us from gaining a better understanding of this fraction of members.

Degree Distributions

The in-degree distribution (distribution of the number of followers) of the empirical Twitter data is shown in fig. 2, while fig. 3 shows the out-degree distribution (distribution of the number of friends). All distributions are plotted using log-log scale. Both the in-degree distribution and the out-degree distribution indicate that a very large fraction of Twitter users have very low number of followers / friends. These correspond to the inactive members, i.e. members who are not interested in creating follow-links with others.

The in-degree distribution shows a power-law decay $p_k \sim k^{-1.0}$ for a significant range of the in-degree below 2000, as shown in fig. 2; but it deviates from power-law for low values of in-degree as well as for very high values of in-degree. This scale-free nature of the in-degree distribution, as observed in our dataset, agrees with results obtained from earlier studies [12] on Twitter.

The out-degree distribution, fig. 3, clearly shows the effect of the restriction on the number of friends: the distribution shows a power-law decay $p_k \sim k^{-1.0}$ for out-degrees below 2000, but a sharp spike is observed at around the degree $k = 2000$, corresponding to an uncharacteristically large fraction of members having near about 2000 friends. This is due to the existence of a significant fraction of members who are unable to increase their number of friends beyond a certain limit near 2000, because they do not have a sufficient number of followers. The same observation is reflected as a discontinuity in the cumulative out-degree distribution - the fraction of members having more than k friends drops abruptly around $k = 2000$, signifying the relatively large fraction of members having out-degree near 2000. The cumulative out-degree distribution for the range [1900, 2100] of out-degree is magnified in fig. 3(b) (inset) to show the discontinuity. To the best of our knowledge, this change in the out-degree distribution of Twitter as a result of the imposed restriction is first being reported in the current study.

Fig. 3(a) also shows an uncharacteristically large fraction of members having 20 (or a few more than 20) friends. This can be explained by the fact that when a new member joins the Twitter network, Twitter recommends a set of 20 existing members for the new member to follow. It is likely that a substantial number of new members choose to follow all of these 20 recommended accounts; again, many of them become inactive without creating any more follow-links (or, after creating a few more follow-links), thus resulting in a relatively large number of members having near about 20 friends. A study of the members having 20 or 21 friends, in our dataset, reveals that a large majority of these accounts have posted less than 10 tweets in their entire lifespan,

thus supporting the claim that they are inactive members.

From the out-degree distribution of the Twitter network given above, it is evident that the topological properties of OSNs can change significantly due to imposed restrictions on node-degree. The primary motivation of the present work is to formulate an analytical framework to study the effects of such restrictions on the degree-distribution of a network, as presented in the next section.

5 Framework for modeling restricted growth dynamics of OSNs

In this section, we develop a framework to model the restricted growth of OSNs in general and Twitter in particular. For this, we need to model the growth dynamics of OSNs (i.e. dynamics of new members joining the network, and creation of friendship-links among members), and then study the effect of the imposed restrictions on the topological properties that emerge due to the growth dynamics.

We model the growth dynamics of OSNs by the *preferential attachment* growth model [8] in which new links are attached preferentially to members who already have a large number of links. The justifications for this choice are as follows. Preferential attachment has been shown to occur in several OSNs [15, 17]. Moreover, preferential attachment is known to produce power-law degree distributions, as is seen in the samples of Twitter network obtained both in our study (detailed in section 4) and in earlier studies on Twitter [12].

In case of directed networks such as Twitter, the preferential attachment model can be divided into two parts: (i) preferential creation of links, where members create new links in proportion to their out-degree, and (ii) preferential reception of links, where members receive new links in proportion to their in-degree. The intuitive explanation for these aspects is that a member who already has many out-links (friends) is socially more active, hence she is more likely to create more out-links; similarly a member who already has many in-links (followers) is a popular member and hence is more likely to get new followers.

We customize the growth model proposed by Krapivsky et. al. [13] (henceforth referred to as the KRR model) which was originally proposed to explain the in-degree and out-degree distributions of the world-wide web using preferential attachment. The process of a web-page having a hyper-link to another is analogous to the process of a user being a follower of another. Just as a well-known web page is more likely to have new web pages linking to it, a popular user with many followers is more likely to be followed by new members. Conversely, just as a web page with many outgoing hyperlinks is more likely to create even more hyperlinks, a socially active member who follows many members is more likely to follow others.

We modify the KRR model by introducing restrictions on the out-degree of nodes, similar to the follow-limit imposed by Twitter. However, the modified model is general enough to be used to model different types of restrictions on node-degree that can be introduced in OSNs, as explained below.

The modified KRR model

In this model, network growth occurs in two distinct steps. At each discrete time step, one of the following events occurs: (i) with probability p, a new node is introduced and it forms a directed out-edge to an existing node, or (ii) with probability $q = 1 - p$, a new directed edge is created between two existing nodes.

The *attachment rate* $A(i, j)$, defined as the probability that a newly-introduced node links to an existing (i, j)-node (i.e. a node of in-degree i and out-degree j), is assumed to be an increasing function of i (preferential creation) but independent of j:

$$A(i, j) = A_i = i + \lambda \tag{1}$$

This is analogous to the intuitive idea that when a new member u joins an OSN (Twitter), she is more likely to form a connection to (follow) a popular member v having many followers, but the number of people followed by v is not likely to influence the choice of u.

The *creation rate* $C(i_1, j_1|i_2, j_2)$, defined as the probability that an edge is created from a (i_1,j_1)-node to a (i_2,j_2)-node, is assumed to be an increasing function of j_1 (preferential creation) and i_2 (preferential reception), but is independent of i_1 and j_2:

$$C(i_1, j_1|i_2, j_2) = C(j_1, i_2) = (i_2 + \lambda)(j_1 + \mu) \tag{2}$$

This again follows the intuition that if u is a socially active member who follows many people already, she is more likely to follow another member v (especially if v is popular herself, having many followers); however, u's decision to follow v is not likely to be influenced by the number of followers of u, nor by the number of people whom v follows.

In the above equations, λ and μ are model parameters that introduce randomness in the preferential attachment rules. They must satisfy the constraints, $\lambda > 0$ and $\mu > -1$, to ensure that the corresponding probabilities are positive for all permissible values of in-degree and out-degree, $i \geq 0$ and $j \geq 1$ (all nodes enter with out-degree 1).

Let $N_{ij}(t)$ denote the average number of nodes in the network, having in-degree i and out-degree j at time t. With the addition of a new node (with probability p) or

a new edge (with probability $q = 1 - p$) at time t, the number N_{ij} may change due to the following events: (i) change in in-degree of $(i-1, j)$-nodes and (i, j)-nodes, and (ii) change in out-degree of $(i, j-1)$-nodes and (i, j)-nodes. These events are discussed individually below.

The number N_{ij} of (i, j)-nodes increases when a new edge is created leading to a $(i-1, j)$-node (of which there are $N_{i-1,j}$ in number); this can happen due to a new node linking to a $(i-1, j)$-node (with probability p) or due to the creation of a new edge leading to a $(i-1, j)$-node (with probability $q = 1 - p$). When the attachment and creation rates are given by equations 1 and 2 respectively, this increase occurs with the rate $(p+q)(i-1+\lambda)N_{i-1,j}$, divided by the normalization factor $\sum_{ij}(i+\lambda)N_{ij} = I + \lambda N$, where $N(t)$ is the total number of nodes in the network at time t, and $I(t)$ is the total in-degree in the network at time t.

On the other hand, the number N_{ij} of (i, j)-nodes gets reduced when a new edge is created leading to a (i, j)-node; this can happen due to a new node linking to a (i, j)-node (with probability p) or due to the creation of a new edge leading to a (i, j)-node (with probability $q = 1 - p$). Hence, this reduction in N_{ij} occurs with the rate $(p+q)(i+\lambda)N_{ij}/(I+\lambda N)$. Since $(p+q) = 1$, therefore the rate of change in $N_{ij}(t)$ due to change in in-degree of nodes is as given in eqn. 3.

$$\frac{dN_{ij}}{dt}_{in} = \left[\frac{(i-1+\lambda)N_{i-1,\,j} - (i+\lambda)N_{ij}}{I+\lambda N}\right] \quad (3)$$

Similarly, there is a gain in N_{ij} when a $(i, j-1)$-node forms a new out-edge (this event occurs with the rate $q(j-1+\mu)N_{i,j-1}/(J+\mu N)$, where $J(t)$ is the total out-degree in the network at time t); and there is a loss in N_{ij} when a (i, j)-node forms a new out-edge (with the rate $q(j+\mu)N_{i,j}/(J+\mu N)$). These events can occur only due to creation of links among existing nodes, hence the rates are multiplied by the probability q. Since the change in out-degree of nodes is restricted due to imposed limits (as in Twitter), the rate of change in $N_{ij}(t)$ due to change in out-degree of nodes is as given in eqn. 4. The terms β_{ij} capture the effects of the restriction; their significance is explained below.

$$\frac{dN_{ij}}{dt}_{out} = q\left[\frac{(j-1+\mu)N_{i,j-1}\beta_{ij} - (j+\mu)N_{ij}\beta_{i,j+1}}{J+\mu N}\right] \quad (4)$$

Thus the total rate of change in the number N_{ij} of (i, j)-nodes is given by eqn. 5.

$$\frac{dN_{ij}}{dt} = \frac{dN_{ij}}{dt}_{in} + \frac{dN_{ij}}{dt}_{out} + p\delta_{i0}\delta_{j1} \quad (5)$$

The last term in eqn. 5 accounts for the introduction of new nodes with in-degree zero and out-degree one, with a probability p at every time-step. δ_{i0} is 1 for $i = 0$ and 1 otherwise; δ_{j1} is 1 for $j = 1$ and 0 otherwise.

Incorporating restrictions in the model

The β_{ij} terms in eqn. 4 capture the effect of the imposed restrictions on the growth dynamics. It is to be noted that since Twitter in particular imposes the restriction only on out-degree of nodes, the β_{ij} factors appear only in eqn. 4. This model can be easily modified to study restrictions imposed on in-degree of nodes (or total-degree, as is done in OSNs like Orkut and Facebook) by including similar β_{ij} terms in eqn. 3.

The role of the β_{ij} terms in eqn. 4 is explained as follows. Due to the imposed restriction (e.g. the Twitter follow-limit), only a fraction of the existing nodes can create new out-links, and β_{ij} is defined such that it equals 1 only for this fraction of nodes. In other words, β_{ij} is defined to be 1 if and only if members having in-degree i are allowed (by the restriction) to have out-degree j.

The β_{ij} terms can be defined according to the restriction that needs to be studied, thus making this model suitable to study restrictions of different types. Let us take the example of the Twitter follow-limit. To generalize the model, let k_c denote the out-degree at which the restriction starts and let the restriction be generalized to an 'α-percent rule' (k_c = 2000 and α = 10 for the real-world Twitter network).

To study version 1 of the Twitter follow-limit (see section 2), β_{ij} is defined as:

$$\beta_{ij} = \begin{cases} 1 & \text{if } j \leq \max\{\, k_c, (1+\frac{1}{\alpha})i \,\}, \forall i \\ 0 & \text{otherwise} \end{cases}$$

Similarly, in order to study version 2 of the Twitter follow-limit (see section 2), β_{ij} can be defined as:

$$\beta_{ij} = \begin{cases} 1 & \text{if } i < k_c \text{ and } j \leq k_c + \frac{1}{\alpha}i \\ 1 & \text{if } i \geq k_c \text{ and } j \leq (1+\frac{1}{\alpha})i \\ 0 & \text{otherwise} \end{cases}$$

As a third example, a 'hard' cut-off at out-degree k_c can be studied using this model simply by defining β_{ij} as

$$\beta_{ij} = \begin{cases} 1 & \text{if } j \leq k_c, \forall i \\ 0 & \text{otherwise} \end{cases}$$

Significance of the model parameters

The model described above has three growth parameters, namely, p (the probability of introduction of a new node), λ and μ (randomness factors in preferential attachment), along with the two parameters α and k_c that are specific to the restriction imposed in Twitter. The significance of α and k_c is obvious from the description of the 'soft' cut-off limit imposed in Twitter. The role of the three growth parameters in modeling the dynamics observed in OSNs in general are as follows.

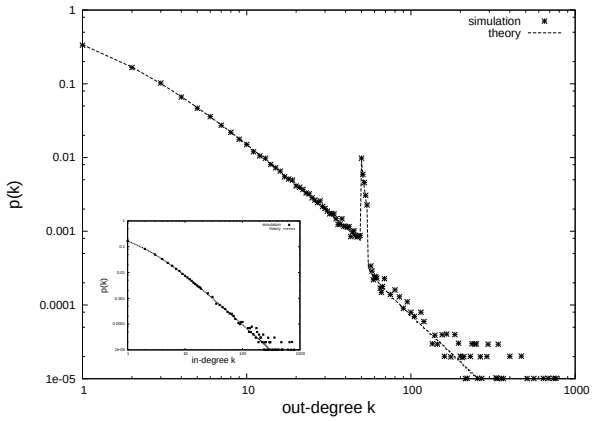

Figure 4: Agreement of simulation and theoretical model for restricted growth of networks (Twitter restriction on out-degree). Parameters: p = 0.01, $\lambda = \mu = 1.0$, k_c = 50, α = 10 (a) out-degree distribution (b) (inset) in-degree distribution

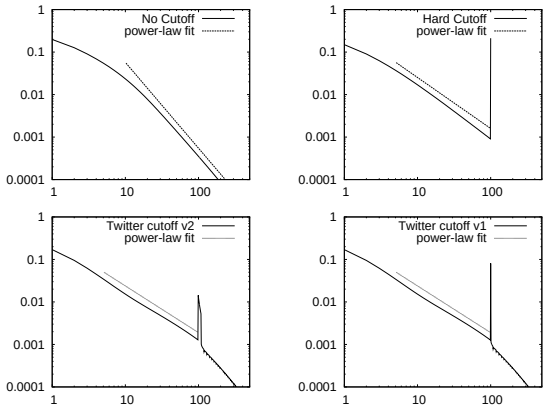

Figure 5: Effect of different forms of restrictions on out-degree distribution (log-log plot). Clockwise from top-left: (a) No restriction, power-law fit with γ = -2.02 (b) A 'hard' cut-off, γ = -1.2 (c) Twitter restriction version 1, γ = -1.1 (d) Twitter restriction version 2, γ = -1.1. Parameters: p = 0.01, $\lambda = \mu$ = 1.0, k_c = 100, α = 10

The parameter p controls the relative number of nodes and edges, i.e. the density of the network. According to the dynamics of the model, the average in-degree and average out-degree are both $1/p$ [13]. A study [10] on Twitter in January 2010 reports that the growth (i.e. new members joining) of Twitter has slowed down considerably in the later half of 2009, but the average number of friends and followers of users have increased. Such effects can be incorporated into the model by tuning the value of p (or even varying it over time).

The parameters λ and μ indicate how closely the dynamics of link-formation in an OSN follow the preferential attachment model (lower values indicate more closeness to preferential attachment). Though the dynamics of several OSNs have been found to be in close agreement with the preferential attachment model [15, 17], estimating these parameters for a real-world OSN is a challenging issue. Moreover, these parameters can change with time in a real-world OSN, e.g. due to the recommendation of selected existing members to new members (as done in Twitter).

6 Results

This section discusses the results obtained using the theoretical model developed in the previous section. Since experiments in the scale of the empirical data collected from Twitter would be too time-consuming, hence the results given are from experiments performed at a much smaller scale (as indicated by the parameter values), however this does not affect the generality of the results.

The parameter p is set to 0.01 throughout all experiments unless otherwise stated, in order to have the average in-degree / out-degree in the same order as that has been recently reported for Twitter [10].

Validation of theoretical model

We validate the theoretical model developed in section 5 by simulating the restricted emergence of the network. The stochastic simulation is continued until the total number of nodes in the network is 10000 and we perform 100 individual realizations and plot the average degree distributions. Eqn. 5 is solved iteratively until the N_{ij} values reach a steady state, and the in-degree (out-degree) distribution is computed as $N_i^{in}(t) = \sum_j N_{ij}(t)$ ($N_j^{out}(t) = \sum_i N_{ij}(t)$). Fig. 4 shows that the agreement between the theory and the simulation results is exact, which validates the correctness of the proposed theoretical framework.

Different types of restrictions

Figs. 5(b,c,d) show the effect of the different forms of restrictions discussed in section 5 on the out-degree distribution, along with the out-degree distribution in the absence of any restriction 5(a). It is evident that 'hard' cut-offs block a much larger fraction of users as compared to the 'soft' cut-off imposed by Twitter, thus justifying their criticism from users of popular OSNs.

Power-law fits to the distributions are also shown in fig. 5. The 'hard' cut-off reduces the absolute value of the power-law exponent (γ) in the out-degree distribution from 2.02 (in absence of cut-off) to 1.2, i.e. the out-degree distribution becomes flatter, as seen in fig. 5(b). Similar reductions in γ have been reported for cut-offs in peer-to-peer networks [11]. The 'soft' cut-offs imposed by Twitter further reduce the absolute value of γ to 1.1 in the region below the cut-off (fig. 5(c) and fig. 5(d)).

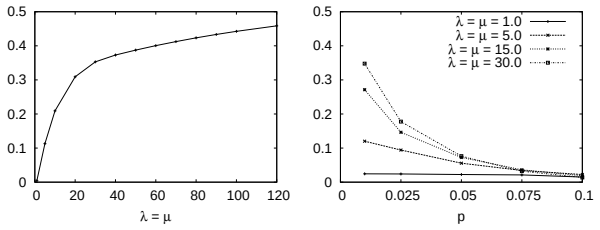

Figure 6: Number of nodes which cross the 'soft' Twitter restriction (version 1), as a fraction of total nodes (a) variation with $\lambda = \mu$ (b) variation with p

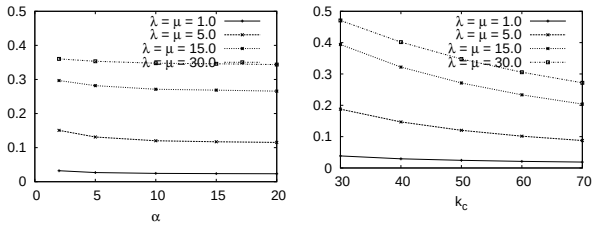

Figure 7: Number of nodes which cross the 'soft' Twitter restriction (version 1), as a fraction of total nodes (a) variation with α (b) variation with k_c

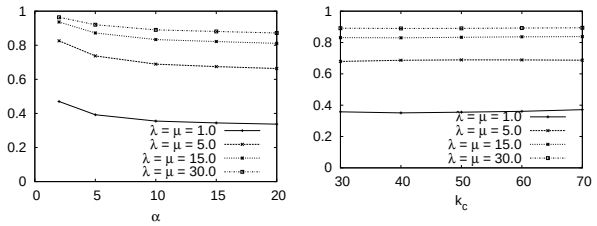

Figure 8: Number of nodes which cross the 'soft' Twitter restriction (version 1), as a fraction of the number of nodes which approach the cut-off (a) variation with α (b) variation with k_c

A smaller absolute value of γ indicates a more homogeneous structure of the network with respect to degrees. This provides scalability to OSNs as messages produced will get equitably distributed among various users, and hence various servers, and would not be directed towards a small group of users (servers).

Also, version 1 of the Twitter cut-off is verified to be a stricter cut-off than version 2, as reflected by the much higher spike in fig. 5(c) compared to that in fig. 5(d) (signifying a larger fraction of nodes that get blocked). It is to be noted that the out-degree distribution predicted by our model for the Twitter restrictions has a power-law coefficient (γ = -1.1) that is very similar to that of the empirical Twitter data (γ = -1.0, as reported in section 4).

Effects of the network dynamics

We study the effects of the network dynamics by measuring the fraction of nodes that can cross the restriction, for various values of the parameters λ, μ and p (whose significance are explained in section 5).

Fig. 6(a) plots the number of nodes which can cross the Twitter cut-off (version 1), as a fraction of total nodes in the network, for different values of $\lambda = \mu$ in the range 1.0 to 120.0 (or a few in this range). Since empirical estimates of these parameters are not available, we have taken $\lambda = \mu$ in all cases, but this can be varied if necessary. The fraction of nodes crossing the cut-off increases rapidly with λ (= μ) for their lower values, but stabilizes for higher values of λ (= μ). This can be explained as follows. For very low values of λ (= μ), the dynamics is almost fully preferential, hence only the very popular members (having high in-degrees) can cross this limit. As λ (= μ) increases, the randomness in the dynamics increases and a larger fraction of nodes can attain in-degrees that enable them to cross the cut-off. This reaches a stability when the system becomes highly random.

Fig. 6(b) plots the fraction of nodes which can cross the Twitter cut-off (version 1) for different values of p in the range 0.01 to 0.1 (or a few in this range). We use different values of $\lambda = \mu$ in the range 1.0 to 30.0 to investigate varying link-creation dynamics ranging from close matches to preferential attachment ($\lambda = \mu$ = 1.0) to more random dynamics ($\lambda = \mu$ = 30.0). As the value of p increases, there is lesser activity (and more growth) in the network, hence a smaller fraction of nodes approach the cut-off; this results in a sharp decay in the fraction of nodes crossing the cut-off, for all cases of $\lambda = \mu$.

Choice of cut-off parameters

The proposed model can also be used to design functions with varying levels of difficulty in overcoming the restriction, as discussed below. Fig. 7 plots the number of nodes which can overcome restrictions similar to the Twitter cut-off (version 1), *as a fraction of the total number of nodes*. Different values of the restriction parameters α (fig. 7(a)) and k_c (fig. 7(b)) are used to experiment with restrictions of different rigidity. Again, we use different values of $\lambda = \mu$ in the range 1.0 to 30.0 to investigate varying link-creation dynamics.

As seen from fig. 7(a), the fraction of nodes overcoming the limit does *not* change appreciably with α for any of the cases. However, for more random dynamics (relatively higher values of $\lambda = \mu$), the fraction of nodes overcoming the limit falls rapidly with increase in k_c (fig. 7(b)). Thus the importance of k_c in the restriction function is to limit the fraction of members in the whole network, that are able to cross an imposed cut-off.

The number of nodes which can overcome restrictions similar to the Twitter cut-off (version 1), *as a fraction of the number of nodes which approach the cut-off*, is plotted in fig. 8. The number of nodes which approach the

cut-off is measured as the sum of the nodes which get blocked at the cut-off and those that cross the cut-off. Interestingly, this fraction of nodes is seen to be relatively invariant with k_c (fig. 8(b)); instead, it reduces with the increase in α, specially in the range $\alpha < 10$ (fig. 8(a)). Hence the parameter α is more effective in deciding what fraction of the members who approach the cut-off are able to overcome it. Moreover, this fraction seems to stabilize for the range $\alpha > 10$; this may be a possible justification of the fact that Twitter uses the value $\alpha = 10$.

Interestingly, the fraction of nodes that are able to overcome the Twitter limit (version 1), for the case λ = μ = 1.0, as predicted by our model (0.03 - 0.04), is in the same order as the corresponding fraction obtained from the empirical data collected from Twitter (0.0668).

Summarizing, the interpretation of the theoretical model points to several interesting results such as (a) cut-offs make the network homogeneous and eases the pressure on hubs, (b) preferentiality hinders users from crossing the restriction, (c) the fraction of users crossing the restriction is almost independent of the parameter α, while (d) the fraction of users stopped by the restriction is independent of the parameter k_c. As such, we feel that this nature of analyses and the interesting observations would be required by the popular OSNs in the recent future, in order to design limits that reduce spam and strain on the system, without affecting legitimate users of the OSN.

7 Conclusion

We summarize the main contributions of the paper. The effects of restrictions on node-degree, on the topological properties of an OSN are studied, taking Twitter as a case-study, and an analytical framework is developed that can be used to study the effects of different forms of such restrictions. We demonstrate how this framework can be used to experiment with different forms of restrictions and growth dynamics and identify suitable values for restriction parameters.

For the sake of simplicity, the model developed does not consider some of the dynamics in the Twitter OSN, such as the recommendation of friends to new users (which may explain the large fraction of users found to have 20 friends), and the convention of 'following-back' that is adopted by many Twitter users. Such factors can be incorporated in future models of Twitter. Also a study that formally relates the degree distributions emerging in OSNs due to the imposed restrictions with the improvement in performance needs to be undertaken. More importantly, this first line of defence (restrictions) needs to be effectively combined with other (anti-spam) techniques to build up a robust spam-filtering system. We plan to pursue these as future work.

8 Acknowledgements

The second author acknowledges LIP6, Paris for sponsoring an internship during which he learnt techniques to crawl Twitter.

References

[1] Twitter. `http://twitter.com/`.

[2] Twitter api wiki / frontpage. `http://apiwiki.twitter.com/`.

[3] Twitter api wiki / rate limiting. `http://apiwiki.twitter.com/Rate-limiting`.

[4] Twitter support: Following limits and best practices. `http://help.twitter.com/forums/10711/entries/68916`.

[5] Twitter blog: Making progress on spam. `http://blog.twitter.com/2008/08/making-progress-on-spam.html`, August 2008.

[6] The 2000 following limit on twitter. `http://twittnotes.com/2009/03/2000-following-limit-on-twitter.html`, March 2009.

[7] Twitter limits explained. `http://www.webtrepreneur.net/twitter-limits-explained/`, September 2009.

[8] BARABASI, A. L., AND ALBERT, R. Emergence of scaling in random networks. *Science* (1999).

[9] CATONE, J. Twitters follow limit makes twitter less useful. `http://www.sitepoint.com/blogs/2008/08/13/twitter-follow-limit-makes-twitter-less-useful/`, August 2008.

[10] GAUDIN, S. Twitter's growth starts losing steam, study finds. `http://www.pcworld.com/article/187349/twitters_growth_starts_losing_steam_study_finds.html`, January 2010.

[11] GUCLU, H., AND YUKSEL, M. Scale-free overlay topologies with hard cutoffs for unstructured peer-to-peer networks. In *IEEE ICDCS '07* (2007), IEEE Computer Society, p. 32.

[12] JAVA, A., SONG, X., FININ, T., AND TSENG, B. Why we twitter: understanding microblogging usage and communities. In *WebKDD / SNA-KDD 2007* (2007), ACM, pp. 56–65.

[13] KRAPIVSKY, P. L., RODGERS, G. J., AND REDNER, S. Degree distributions of growing networks. *Phys. Rev. Lett. 86*, 23 (Jun 2001), 5401–5404.

[14] KRISHNAMURTHY, B., GILL, P., AND ARLITT, M. A few chirps about twitter. In *WOSN '08* (2008), ACM, pp. 19–24.

[15] KUMAR, R., NOVAK, J., AND TOMKINS, A. Structure and evolution of online social networks. In *KDD* (2006), ACM, pp. 611–617.

[16] LATAPY, M., AND MAGNIEN, C. Complex network measurements: Estimating the relevance of observed properties. In *INFOCOM* (2008), pp. 1660–1668.

[17] MISLOVE, A., KOPPULA, H. S., GUMMADI, K. P., DRUSCHEL, P., AND BHATTACHARJEE, B. Growth of the flickr social network. In *WOSN '08* (2008), pp. 25–30.

[18] MITRA, B., DUBEY, A., GHOSE, S., AND GANGULY, N. How do superpeer networks emerge? In *IEEE INFOCOM* (2010).

[19] MOORE, R. J. Twitter data analysis: An investor's perspective. `http://www.techcrunch.com/2009/10/05/twitter-data-analysis-an-investors-perspective/`, October 2009.

[20] OWYANG, J. The many challenges of social network sites. `http://www.web-strategist.com/blog/2008/02/11/the-many-challenges-of-social-networks/`, February 2008.