Assessing the Effects of a Soft Cut-Off in the Twitter Social Network

Saptarshi Ghosh, Ajitesh Srivastava, Niloy Ganguly

Assessing the Effects of a Soft Cut-off in the Twitter Social Network

Saptarshi Ghosh[1]*, Ajitesh Srivastava[2], and Niloy Ganguly[1]

[1] Department of CSE, IIT Kharagpur, India
[2] Department of CSIS, BITS Pilani, India

Abstract. Most popular OSNs currently restrict the number of social links that a user can have, in order to deal with the problems of increasing spam and scalability in the face of a rapid rise in the number of users in recent years. However such restrictions are often being criticized by socially active and popular users, hence the OSN authorities are facing serious design-choices while imposing restrictions; this is evident from the innovative 'soft' cut-off recently imposed in Twitter instead of the traditional 'hard' cut-offs in other OSNs. Our goal in this paper is to develop an analytical framework taking the restriction in Twitter as a case-study, that can be used to make proper design-choices considering the conflicting objectives of reducing system-load and minimizing user-dissatisfaction. We consequently define a simple utility function considering the above two objectives, and find that Twitter's policy well balances both. From a network science perspective, this is the first analysis of 'soft' cut-offs in any sort of network, to the best of our knowledge.

Keywords: Online social network, Twitter, soft cut-off, restricted network growth, utility function for restrictions

1 Introduction

Online Social Networks (OSNs) have experienced an exponential rise in the number and activity of users in recent years. As a result, these OSNs are frequently facing scalability issues such as high latency and increased down-time [17] which lead to discontent among users. The situation is aggravated by spammers who typically establish social links with thousands of users and then use the methods of communication provided to disseminate spam. Several popular OSNs have adopted a common 'tool' to deal with these issues: they have imposed a limit or cut-off on the number of friends/social links that a user can have (i.e. on the node-degree), e.g. 1000 in Orkut and 5000 in Facebook. Such limits help in reducing the load on the OSN infrastructure - since most OSNs support real-time one-to-all-friends communications, controlling the number of friends of users is an effective way to reduce message overhead. Moreover, these restrictions also prevent spammers from indiscriminately increasing their links.

Twitter (*www.twitter.com*), one of the OSNs worst affected by the above problems, has placed a more intelligent 'soft' cut-off [1] on the number of links

* Correspondence to: Saptarshi Ghosh, Department of CSE, IIT Kharagpur, Kharagpur - 721302, India. Email: saptarshi.ghosh@gmail.com

a user can create. The Twitter social network is a directed network where an edge $u \rightarrow v$ implies that user u 'follows' user v i.e. u has subscribed to receive all messages posted by v. In Twitter terminology, u is a 'follower' of v and v is a 'following' of u. The out-degree (number of followings) of u is thus a measure of u's social activity or her interest to collect information from other users. Analogously, the in-degree of u (number of followers who are interested in u's posts) is a measure of u's popularity in Twitter.

The growing popularity of Twitter in recent years has not only led to high system-load due to increasing user-activity, but also to high levels of "Follow Spam" [2] where spammers indiscriminately follow numerous users, hoping to get followed back. To reduce strain on the website [1] and control follow spam, Twitter enforced a restriction on the number of people that a user can follow (i.e. on the out-degree), in August 2008 [2]. 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 stated in the Twitter Support webpages [1]. However, Twitter does not specify the restriction fully in public [2] (security through obscurity). This has led to several conjectures regarding the Twitter follow-limit; among these, the most widely believed one, known as the "10% rule" [3], is as follows. If a user u has u_{in} number of followers (in-degree), then the maximum number of users whom u can herself follow (maximum possible out-degree) is $u_{out}^{max} = \max\{2000,\ 1.1 \cdot u_{in}\}$.

However, restrictions on the number of links are presently being frequently criticised by the socially active and popular legitimate users of OSNs, as an encroachment on their freedom to have more friends [6]. In fact, the 'soft' cut-off in Twitter is the first attempt by an OSN towards designing restrictions that adapt to the requirements of popular legitimate users (unlike the 'hard' cut-offs in Facebook/Orkut), and hence aim to minimize user-dissatisfaction along with fulfilling other objectives (e.g. reducing system-load).

Evidently, the OSN authorities today are facing several design-choices while designing restrictions, such as - at what degree should the restriction be imposed so that a desired reduction in the system-load can be achieved without affecting a large number of legitimate users? In order to explore and utilise the full potential of restrictions on node-degree, an analytical model that helps to make such design-choices, rather than ad-hoc engineering solutions, has become a necessity. The goal of this paper is to formulate such a model using the methods of network science, taking the Twitter follow-limit as a case study.

Restrictions on node-degree have significant effects on the topology of present-day OSNs, as was first observed by us for the Twitter network in [8]. In this paper, we extend the rudimentary model proposed in [8] to develop a complete analytical framework that can be used to predict the emerging *degree distribution* of an OSN in the presence of different forms of restrictions. We demonstrate the effectivity of enumerating the degree distribution (for restricted growth) by our model by formulating a simple utility function for restrictions, whose optimization would enable the OSN authorities to design restrictions that suitably

balance the two conflicting objectives of reducing system-load and minimizing dissatisfaction among users.

There have been several studies on the topological characteristics that emerge as a result of various growth dynamics in OSNs [5, 13, 15]; however, to the best of our knowledge, ours is the first set of work on analysing the effects of restrictions on node-degree on these dynamics. From a network science perspective, though there have been studies on the effects of 'hard' cut-offs on node-degree (e.g. in peer-to-peer networks [9, 16]), there has not been any prior analysis on network-growth in the presence of 'soft' cut-offs (as has been imposed in Twitter), according to our knowledge.

The rest of the paper is organized as follows. Section 2 describes the effects of the Twitter restriction on the topology of the OSN. The analytical framework for modeling network growth in the presence of restrictions is developed in Sect. 3 while the insights drawn using the model are discussed in Sect. 4. Conclusions from the study are drawn in Sect. 5.

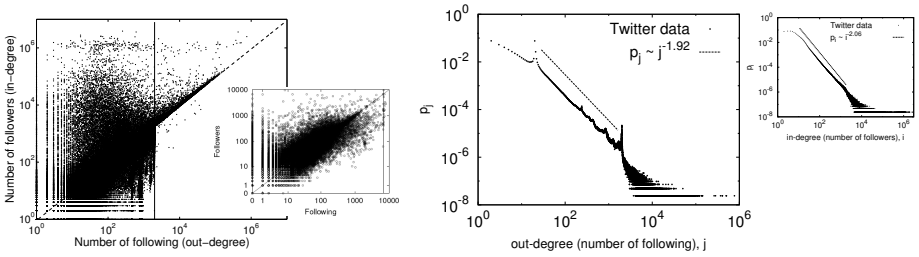
2 Empirical Measurements on the Twitter Social Network

The Twitter OSN has been of interest to researchers since 2007 and there have been several attempts [8, 10, 12, 14] to crawl the Twitter network [3]. Recently a large crawl of the entire Twitter social network in July 2009, containing about 41.7 million nodes and 1.47 billion follow-edges, has been made publicly available [14]; we use this data for empirical measurements in this paper. In this section, we discuss the statistics of followers (in-degree) and followings (out-degree) of users in the Twitter social network, which clearly shows the effects of the restriction on the network topology.

Scatter plot: Fig. 1a compares the scatter plot of the followers-followings spread in Twitter as in July 2009, with the corresponding scatter plot in February 2008 which was before the restriction was imposed (reproduced from [12] as an inset). While the scatter plot in 2008 is almost symmetrical about $x = y$, the scatter plot in 2009 has a sharp edge at $x = 2000$ due to the restriction at this degree. Users having more than 2000 followings (out-degree x) now need to have a sufficient number of followers (in-degree y), such that their out-degree remains less than 110% of their in-degree (i.e. they lie to the left of the $x = 1.1\,y$ line); this verifies the '10% rule' stated in Sect. 1. Note that there exists a small fraction of users who violate the 10% rule; possibly Twitter relaxes its restriction for some users, such as those who joined the OSN before the restriction was imposed.

Degree Distributions: The in-degree and out-degree distributions of the Twitter OSN, as in July 2009, are shown in Fig. 1b. The in-degree distribution (inset) shows a power-law decay $p_i \sim i^{-2.06}$ over a large range of in-degrees (the power-law exponents are estimated by the method in [7]); however, the out-degree distribution clearly shows a departure from the power-law nature that was observed by measurements on Twitter *before* the restriction was imposed [10, 12].

[3] We ourselves crawled 1 million users during Oct-Nov 2009; though these data exhibited the effects of the restriction on the network properties, as observed in [8], it suffered from the known bias of partial BFS-sampling towards high-degree nodes.



(a) (b)

Fig. 1. (a) Scatter plot of followers-followings spread in Twitter: main figure - in July 2009 (along with the lines $x = 1.1y$ and $x = 2000$), inset - in Jan-Feb 2008 (reproduced from [12]) (b) Degree distributions of Twitter OSN as in July 2009: main plot - out-degree distribution, inset - in-degree distribution

Now, the power-law $p_j \sim j^{-1.92}$ for the out-degrees below the point of restriction is followed by a sharp spike at around out-degree $j = 2000$, and a rapid decay in the distribution beyond this point. This is because a significant number of users are unable to increase their out-degree beyond a certain limit near 2000 as they do not have sufficient in-degree (followers). The out-degree distribution also shows a peak at $x = 20$ because till 2009, Twitter used to recommend an initial set of 20 people for every newcomer to follow by a single click, and many newcomers took up this offer (as also observed in [14]).

3 Modeling Restricted Growth Dynamics of OSNs

In this section, we extend the model we proposed in [8] to develop a complete analytical framework for modeling the growth of OSNs in general and Twitter in particular. We model the growth dynamics in an OSN (i.e. joining of new users, creation of new social links) by the *preferential attachment* model [4] which has been experimentally shown to occur in several OSNs [13, 15]. Also, it produces power-law degree distributions similar to the empirical distributions in Twitter before the restriction was imposed [10, 12].

Our proposed model is a customized version of the network-growth model proposed by Krapivsky et. al. [11] (henceforth referred to as the KRR model), which we modify by introducing restrictions on out-degree, similar to the follow-limit imposed in Twitter. We first briefly discuss the modification introduced by us in [8] for the sake of completeness.

3.1 The Model Proposed in [8]

In this model, any one of the following events occurs at each discrete time-step: (1) with probability p, a new node is introduced and it forms a directed out-edge to an existing node, or (2) with probability $q = 1 - p$, a new directed edge is created between two existing nodes.

The probability that a new node (event 1) links to an (i, j)-node (i.e. a node having in-degree i and out-degree j) is assumed to be proportional to $(i + \lambda)$, since intuitively a new user is more likely to link to (follow) a popular user having many followers (high in-degree). Analogously, the probability that a new edge (event 2) is created from a $(i_1,\ j_1)$-node to a $(i_2,\ j_2)$-node is assumed to

be proportional to $(i_2 + \lambda)(j_1 + \mu)$. Here λ and μ are model parameters that introduce randomness in the preferential attachment rules [11]. Let $N_{ij}(t)$ be the average number of (i, j)-nodes in the network at time t. The model considers the following *rate-equations* to track how N_{ij} changes with time.

Change in N_{ij} due to change in out-degree of nodes: Restrictions on out-degree are incorporated in the model by introducing the β_{ij} terms in (1) below, where β_{ij} is defined to be 1 if users having in-degree i are allowed (by the restriction) to have out-degree j, 0 otherwise. N_{ij} increases when a $(i, j-1)$-node forms a new out-edge (event 2); however, only those $(i, j-1)$-nodes are allowed to do this for whom $\beta_{ij} = 1$. This event occurs with the rate $q(j-1+\mu)N_{i,j-1}\beta_{ij}$ divided by the normalization factor $\sum_{ij}(j+\mu)N_{ij}\beta_{i,j+1}$. Similarly, N_{ij} gets reduced when an (i, j)-node (having $\beta_{i,j+1} = 1$) forms a new out-edge (event 2). Thus the rate of change in $N_{ij}(t)$ due to change in out-degree of nodes is:

$$\frac{dN_{ij}}{dt}\Big|_{out} = q \cdot \frac{(j-1+\mu)N_{i,j-1}\beta_{ij} - (j+\mu)N_{ij}\beta_{i,j+1}}{\sum_{ij}(j+\mu)N_{ij}\beta_{i,j+1}} \tag{1}$$

Change in N_{ij} due to change in in-degree of nodes: This case is similar to the above case, only we are not considering any restriction on in-degrees.

$$\frac{dN_{ij}}{dt}\Big|_{in} = \frac{(i-1+\lambda)N_{i-1,\,j} - (i+\lambda)N_{ij}}{\sum_{ij}(i+\lambda)N_{ij}} \tag{2}$$

Hence the total rate of change in $N_{ij}(t)$ is given by

$$\frac{dN_{ij}}{dt} = \frac{dN_{ij}}{dt}\Big|_{in} + \frac{dN_{ij}}{dt}\Big|_{out} + p\,\delta_{i0}\,\delta_{j1} \tag{3}$$

where the last term accounts for the introduction of new nodes with in-degree 0 and out-degree 1 (Kronecker's delta function δ_{xy} is 1 for $x = y$ and 0 otherwise).

This model can be used to study various restrictions by suitably defining the β_{ij} terms in (1). To study the Twitter follow-limit, we define β_{ij} for a generalized 'κ-% rule' starting at out-degree s ($\kappa = 10$ and $s = 2000$ in Twitter, see Sect. 1) as

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

3.2 Extending the Model to Find Degree Distributions Analytically

The preliminary model in [8] is extended by solving (3) to analytically compute the emerging degree distributions in presence of 'soft' cut-offs. We demonstrate the solution for the commonly believed version of the Twitter restriction, other variations of 'soft' cut-offs can be analysed by a similar technique.

At time t, let $N(t)$ be the total number of nodes in the network, and let $I(t)$ and $J(t)$ be the total in-degree and total out-degree respectively. Since at every time-step, a new edge is added and a new node is added with probability p,

$$N(t) = \sum_{ij} N_{ij} = p\,t, \quad I(t) = \sum_{ij} iN_{ij} = J(t) = \sum_{ij} jN_{ij} = t \tag{4}$$

Thus parameter p controls the relative number of nodes and edges in the network. The denominator (normalizing factor) in (2) equals $(I + \lambda N)$. For the denominator in (1), we make a simplifying approximation - we assume that at a given time,

the number of nodes that are actually blocked by the restriction from increasing their out-degree (i.e. number of (i,j)-nodes for which $\beta_{i,j+1}$ is 0) is negligibly small compared to the total number of nodes in the network, which implies

$$\sum_{ij}(j+\mu)N_{ij}\beta_{i,j+1} \simeq \sum_{ij}(j+\mu)N_{ij} = (J+\mu N) \tag{5}$$

Note that this approximation is valid only for large values of μ, when the fraction of nodes blocked by the restriction actually becomes very small (see Sect. 4.2). By solving (3) with the above approximation for few small values of i,j, it is seen that $N_{ij}(t)$ grows linearly with time [11]; hence we can substitute

$$N_{ij}(t) = n_{ij}\,t \tag{6}$$

where n_{ij} is the (constant w.r.t. time) rate of increase in the number of (i,j)-nodes. Substituting (4) and (6) in (3) gives a recursion relation for n_{ij}:

$$n_{ij} = \frac{(i-1+\lambda)n_{i-1,\,j} - (i+\lambda)n_{ij}}{1+\lambda p} + \frac{q(j-1+\mu)n_{i,j-1}\beta_{ij} - q(j+\mu)n_{ij}\beta_{i,j+1}}{1+\mu p} + p\delta_{i0}\delta_{j1} \tag{7}$$

For brevity, we denote the first fraction on the right-hand side in (7) as A_{ij}.

To simplify the computation of the functional form of the degree distribution, we assume (as in the original KRR model [11]) that the power-law exponents of the in-degree and out-degree distributions are equal, which implies $\lambda = (\mu+1)/q$. The exponents were actually found [10] to be equal for the Twitter OSN before the restriction was imposed. Since we are studying restrictions only on out-degree (as in Twitter), we shall henceforth consider only the out-degree distribution. The in-degree distribution can be computed by the original KRR model [11] and will be of the form of a power-law for the entire range of in-degrees.

Let $N_j^{out}(t) = \sum_i N_{ij}(t)$ be the number of nodes with out-degree j at time t; using (6), $N_j^{out}(t) = t\sum_i n_{ij} = tg_j$, where $g_j = \sum_i n_{ij}$. Thus the out-degree distribution at j (i.e. fraction of nodes with out-degree j) can be obtained as $N_j^{out}(t)/N(t) = g_j/p$. To obtain the complete out-degree distribution, we solve (7) to get $g_j = \sum_i n_{ij}$ for all j by considering the following cases.

Case 1: *$j<s$ (before the starting point of cutoff):* Since there is no restriction for $j < s$, the model behaves similar to the original KRR model [11]; hence

$$g_j = G\cdot\frac{\Gamma(j+\mu)}{\Gamma(j+1+q^{-1}+\mu q^{-1})} \sim j^{-(1+q^{-1}+\mu p q^{-1})} \tag{8}$$

where $\Gamma()$ is the Euler gamma function, and G is a constant. Note that (8) is actually an approximation under assumption (5); in reality, the out-degree distribution for $j<s$ is also slightly affected by the restriction (see Sect. 4.1).

Case 2: *$j=s$ (at the starting point of the cutoff):* Let α denote the fraction $\frac{1}{(1+1/\kappa)}$ in case of a κ-% rule ($\kappa = 10$ in Twitter). A node can have an out-degree $j>s$ only if it has an in-degree $i \geq \alpha j$, implying that for $\beta_{i,j+1}$(for $j \geq s$) to be 1, $i \geq \alpha(j+1)$. Hence, for $j = s$, (7) becomes

$$n_{is} = \begin{cases} A_{is} + \frac{q(s-1+\mu)n_{i,s-1}}{1+\mu p} & i < \alpha(s+1) \\ A_{is} + \frac{q(s-1+\mu)n_{i,s-1} - q(s+\mu)n_{is}}{1+\mu p} & i \geq \alpha(s+1) \end{cases} \tag{9}$$

We use a standard technique [11] to solve rate equations: summing (9) for all $i \geq 0$, the terms in A_{is} disappear (they cancel out each other, except the first term in the first equation, i.e. for the case $i = 0$, but that term is zero), and we get

$$g_s = \frac{s-1+\mu}{s+(1+\mu)q^{-1}} \cdot g_{s-1} \quad + \quad \frac{s+\mu}{s+(1+\mu)q^{-1}} \cdot c_s \tag{10}$$

where g_{s-1} can be computed by (8) and $c_s = (\sum_{i=0}^{\lfloor \alpha(s+1) \rfloor} n_{is})$ is the rate of increase in the number of nodes that have out-degree s but cannot increase their out-degree further (i.e. (i, j)-nodes for which $j = s$ and $\beta_{i,s+1} = 0$). Let $\lfloor \alpha(s+1) \rfloor$ be denoted by d. To compute c_s, we sum (9) in the range $0 \leq i \leq d$ to get

$$c_s = \frac{1}{1+\lambda p} \cdot \Big[(s-1+\mu) \sum_{i=0}^{d} n_{i,s-1} \; - \; (d+\lambda) n_{ds} \Big] \tag{11}$$

where n_{ds} can be obtained as

$$n_{ds} = \frac{(s+\mu-1)\Gamma(d+\lambda)}{\Gamma(d+\lambda(1+p)+2)} \sum_{k=0}^{d} \frac{\Gamma(k+\lambda(1+p)+1)}{\Gamma(k+\lambda)} \cdot n_{k,s-1} \tag{12}$$

from (9) after some algebraic manipulations (omitted for sake of brevity). The terms $n_{i,s-1}$ in (11) and (12) can be evaluated from the original KRR model (eqn. 18 in [11]) since they are not affected by the restriction starting from $j = s$. Substituting c_s from (11) into (10), we can obtain a closed-form expression for the degree distribution g_s/p at $j = s$. Equations (10) and (11) can be used to estimate the fraction of members blocked at the point of cut-off, as detailed in Sect.4.

Case 3: *$j > s$ (beyond the starting point of cutoff):* In this region, (7) becomes

$$n_{ij} = \begin{cases} 0 & i < \alpha j \\ A_{ij} + \frac{q(j-1+\mu)n_{i,j-1}}{1+\mu p} & \alpha j \leq i < \alpha(j+1) \\ A_{ij} + \frac{q(j-1+\mu)n_{i,j-1} - q(j+\mu)n_{ij}}{1+\mu p} & i \geq \alpha(j+1) \end{cases} \tag{13}$$

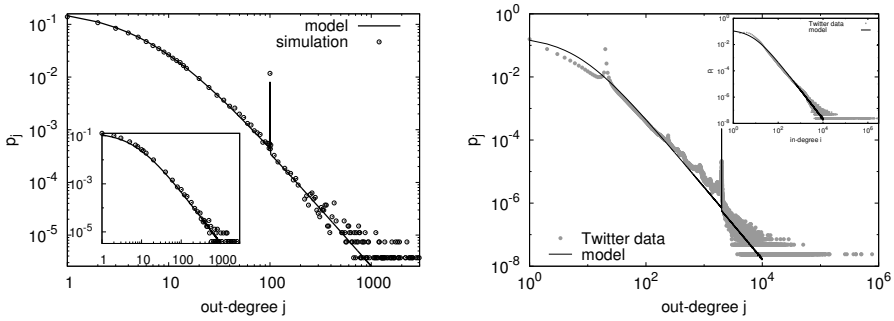
since nodes having in-degree $i < \alpha j$ cannot have out-degree $j (> s)$, nodes having in-degree $\alpha j \leq i < \alpha(j+1)$ can have out-degree j but not $j+1$, and nodes with in-degree $i \geq \alpha(j+1)$ can increase their out-degree from j to $j+1$. Proceeding similarly as in the case $j = s$, and adding (13) over all $i \geq 0$, we get

$$g_j = \frac{j-1+\mu}{j+(1+\mu)q^{-1}} \cdot [g_{j-1} - c_{j-1}] + \frac{j+\mu}{j+(1+\mu)q^{-1}} \cdot c_j \tag{14}$$

where $c_j = \sum_{i=0}^{\lfloor \alpha(j+1) \rfloor} n_{ij}$ is the rate of increase in the number of nodes which have out-degree j but cannot increase their out-degree further, unless their in-degree increases. Proceeding from (14) in a similar way as in the case $j = s$, we can derive analytical expressions for g_j and c_j for $j > s$ iteratively using the values of g_{j-1} and c_{j-1} (e.g. g_{s+1} and c_{s+1} can be derived using g_s and c_s and so on). Details are being omitted for brevity.

3.3 Values of Model Parameters Used for Experiments

The parameter p (ratio of nodes to edges in the network) is set to 0.028 as measured from the empirical data described in Sect.2. Estimating λ and μ (which



(a) (b)

Fig. 2. (a) Agreement of simulation and proposed model (b) Fitting empirical Twitter data with model (main plots: out-degree distributions, inset: in-degree distributions)

indicate the level of randomness in link-creation dynamics) for an OSN is a challenging issue; moreover, they can change with time e.g. due to the recommendation of popular users to others in Twitter. Hence we conduct experiments for different values of these parameters. Since the model assumes $\lambda = (\mu{+}1)/(1{-}p)$, we report results for different values of μ only. Parameters of the restriction function are set to $\kappa{=}10$ and $s{=}2000$ (as in Twitter) unless otherwise stated.

3.4 Validating Proposed Model with Simulated and Empirical Data

Correctness of the proposed model is validated by simulating the restricted growth of the network. Since experiments in the scale of the empirical Twitter data are infeasible, simulations were performed for 100,000 nodes and cut-off $s = 100$ (Fig. 2a). Though the model gives approximate solutions for low values of μ (as stated in Sect. 3.2), Fig. 2a shows almost exact agreement between the theory and simulation for $\mu = 6.0$ (this value fits the empirical distributions for Twitter). Exact agreement was obtained in our experiments for $\mu > 50.0$ (results not reported for brevity).

The empirical in-degree and out-degree distributions of Twitter (described in Sect. 2) show excellent fit with those obtained from the model using $\mu = 6.0$ (Fig. 2b). This signifies that the proposed model successfully captures the growth dynamics of the Twitter OSN. However, the empirical out-degree distribution deviates from the theoretical one in two aspects: (i) the empirical distribution has a peak at out-degree 20, which is explained in Sect. 2, and (ii) the spike at out-degree $s = 2000$ is lower in the empirical data as compared to that in the theory; this can be explained by the following two factors. First, there exist a few thousand users in the empirical data who violate the 10% rule, as stated in Sect. 2. Second, we have observed that many Twitter users who actually get blocked by the restriction reduce their out-degree by *un-following* some of their current followings; this naturally leads to a smaller spike at s and a corresponding rise in the fraction of users having out-degree a little less than s.

4 Insights from the Model

Now that the model is validated and is able to reproduce the degree distributions of the Twitter OSN, we use the model to draw various insights on 'soft' cut-offs.

4.1 Effects of Restrictions on Degree Distributions

'Hard' cut-offs in peer-to-peer networks are known to cause a reduction in the absolute value of the power-law exponent γ of the degree distribution below the cut-off degree [9]. Our experiments [8] show a similar effect on the exponent $|\gamma_{out}|$ of the out-degree distribution due to 'soft' cut-offs in directed networks like Twitter; this can be explained by re-considering our approximation in (5). The denominator in (1), which needs to be evaluated for only those nodes that are currently *not* blocked by the restriction (i.e. (i,j)-nodes for which $\beta_{i,j+1} = 1$), is in fact

$$\sum_j \Big[(j+\mu)\sum_i N_{ij}\beta_{i,j+1}\Big] = \sum_j \Big[(j+\mu)\sum_{i\geq\lceil\alpha(j+1)\rceil} n_{ij}t\Big] = (1+\mu p)t - \zeta t \quad (15)$$

where $\zeta = \sum_j (j+\mu)c_j$ is the unknown term. Thus the denominator of the second fraction on the right-hand side in (7) should actually be $(1+\mu p-\zeta)$. Proceeding as in Sect. 3, it can be shown that in the range $j<s$, $|\gamma_{out}|$ reduces from $(1+q^{-1}+\mu pq^{-1})$ in absence of any restriction (as stated in (8)) to $(1+(1-\zeta)q^{-1}+\mu pq^{-1})$ in presence of the 'soft' cut-off modeled in Sect. 3.

A smaller $|\gamma|$ indicates a more homogeneous structure of the network with respect to node-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). The theoretical reduction in $|\gamma_{out}|$ is also validated from real data of the Twitter OSN where $|\gamma_{out}|$ has decreased after the imposition of the cut-off, from 2.412 as reported in [10] to 1.92 in the data described in Sect. 2.

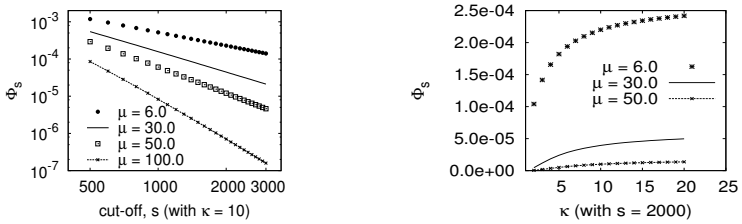
4.2 Quantifying the Fraction of Users Blocked due to the Restriction

In absence of any restriction, g_j decays as $g_j = (j-1+\mu)g_{j-1}/(j+(1+\mu)q^{-1})$ [11]. Comparing this with (10), we see that due to the 'soft' cut-off at $j=s$, the fraction of nodes having out-degree s (i.e. g_s/p) includes the following additional term, which accounts for the spike in the out-degree distribution at this point:

$$\phi_s \;=\; \frac{s+\mu}{s+(1+\mu)q^{-1}} \cdot \frac{c_s}{p} \quad (16)$$

where c_s is obtained from (11). For $s \gg \mu$ and $q \simeq 1$ (for a real-world OSN, typically cut-off s is large and $p = 1-q$ is very small), $\phi_s \simeq c_s/p$ which is an estimate of the fraction of nodes (users) blocked at the point of cut-off. The effects of different parameters on ϕ_s are discussed below.

Our experiments indicate that ϕ_s approximately varies as inversely proportional to the network density p (graphs not shown for lack of space), since for higher p (i.e. when joining of new users dominates link-creation by existing users), the number of nodes reaching the cut-off gets reduced. The network density of OSNs is known to vary non-monotonically over time [13]; hence in practice, parameters of the restriction function (e.g. s and κ) may be varied depending on the dynamics of the network at different stages. ϕ_s also reduces rapidly with increase in the randomness parameter μ (graphs not shown) - for more random dynamics, new links get distributed among a large number of nodes, resulting in a smaller fraction of nodes approaching the cut-off.



(a) (b)

Fig. 3. Variation of the fraction of users blocked at $j = s$ (i.e. height of spike in out-degree distribution) (a) with s (log-log plot) (b) with κ ($p = 0.028$, $\mu = 6.0$)

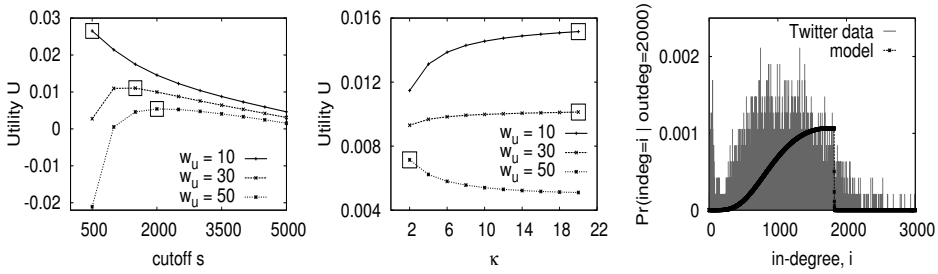
Figures 3a and 3b show the variation in ϕ_s with the restriction parameters s and κ respectively; we use different values of μ to investigate varying link-creation dynamics (from highly preferential to more random). ϕ_s shows a power-law decay with increasing s (Fig. 3a in log-log scale); for lower values of s, a larger fraction of users get blocked leading to a greater reduction in the system-load, but at the risk of increased user-dissatisfaction. Similarly, with increase in κ, a higher in-degree becomes necessary to cross the cut-off resulting in a larger fraction of blocked users; as shown in Fig. 3b, ϕ_s has a parabolic increase with κ.

4.3 Using the Proposed Framework to Design Restrictions

A restriction imposed in an OSN can be said to be effective only if it achieves both the conflicting objectives - a desired reduction in system-load and minimizing dissatisfaction among blocked users. Our proposed model can be used in the process of designing effective restrictions as demonstrated below.

We define a utility function for a restriction as $U = L - w_u B$ where L is the reduction in the number of links due to the restriction (an estimate of reduction in system-load caused by message communication along social links) and B is the fraction of blocked (dissatisfied) users; w_u is the relative weight given to the objective of minimizing user-dissatisfaction and can be chosen suitably by design engineers. For a restriction at out-degree s, we compute $L = (\sum_{j \geq s} j g_j^0 - \sum_{j \geq s} j g_j)$ where g_j is as obtained in Sect. 3 in presence of the restriction while g_j^0, the corresponding quantity in an *unrestricted* network, is computed using the original KRR model (see (8)). Note that our model assumes $g_j = g_j^0$ for $j < s$ as stated in Sect. 3. As discussed above, B can be approximated as $\phi_s \simeq c_s/p$.

Figure 4a shows the variation in utility U with s for a $\kappa = 10\%$ soft cut-off, for different w_u. In each case, the maximum value of U attained is marked. For low w_u, when much higher emphasis is laid on reducing system-load, a low cut-off degree is the best choice. However, as w_u increases, low values of s reduce U since a large fraction of users gets blocked; hence the optimal s occur at higher values. Interestingly, the optimal value for s in the case $w_u = 50$ matches with the value of 2000 chosen in Twitter. The variation in U with κ (for fixed $s = 2000$) is shown in Fig. 4b. For low w_u (higher emphasis on reducing system-load), U increases with κ as more users gets blocked from creating new links; on the contrary, U decreases with κ for higher w_u. It is seen that for $w_u = 50$, the decrease in U stabilizes around the value $\kappa = 10$ that matches with the chosen



(a) (b) (c)

Fig. 4. (a) Variation of utility U with cut-off s (b) Variation of U with κ (c) Comparing in-degree distribution of nodes having out-degree 2000 according to empirical Twitter data (shown in grey) and model (shown in black) (for all plots, $p = 0.028$, $\mu = 6.0$)

value in Twitter. Such analyses are an efficient way for the OSN authorities to make design-choices while imposing restrictions, so that both the objectives of reducing system-load and minimizing user-dissatisfaction can be balanced.

4.4 Estimating the Population of Spammers in the OSN

The population of spammers in an OSN like Twitter can be roughly estimated from the in-degree distribution of users who get blocked at the cut-off. Since the in-degree and out-degree of most *legitimate* users in Twitter are highly correlated [10], among the users blocked at the cut-off, the legitimate ones can be expected to have relatively high in-degrees (number of followers); on the contrary, spammers are likely to have very low in-degrees even when their out-degrees reach the cut-off. According to the model, the number of (i, s)-nodes ($i < \alpha(s+1)$ for nodes blocked at s) at time t is $N_{is}(t) = n_{is}t$, where n_{is} can be computed from (12) by substituting i for d. Since the number of nodes having out-degree s at time t is $g_s t$ (as computed in Sect. 3), n_{is}/g_s gives the value of the said in-degree distribution (conditional to having out-degree s) at in-degree i.

Figure 4c compares the in-degree distribution of nodes having out-degree $s = 2000$, as obtained from the model (for $\mu = 6.0$) and that from the empirical Twitter data. The sharp drop in the theoretical distributions occurs at the minimum in-degree 1820 required to overcome the restriction. Since the model does not consider follow-spammers, most nodes having out-degree s have relatively high in-degrees (corresponding to legitimate users) in the theoretical distribution. In contrast, the Twitter data contains a much higher fraction of 'follow spammers' having low in-degrees and out-degree 2000.

5 Conclusion

In this paper, we take the first step towards analysing restrictions on node-degree in OSNs as well as in the modeling of 'soft' cut-offs in any type of network. We analyse the dependence of the fraction of blocked users on the restriction parameters, such as a power-law reduction with the cut-off degree s and a parabolic increase with κ. We also propose a utility function for restrictions, that helps to balance the conflicting objectives of reducing system-load and minimizing

user-dissatisfaction; this gives practical insights on the choice of values for the restriction parameters, and justifies the choices made in Twitter. Such analyses will be essential to OSN-authorities in recent future for systematically designing restrictions that meet their goals.

Soft cut-offs can be expected to become the chosen type of restriction in all types of OSNs in recent future instead of the frequently criticized 'hard' cut-offs, as they can be easily tuned to adjust to the demands of different types of users. Soft cut-offs can also be applied in *undirected* OSNs (e.g. Facebook, Orkut) by differentiating between the initiator of a social link and the acceptor, and users can be restricted from initiating arbitrary number of links.

References

1. Twitter help center: Following rules and best practices. http://support.twitter.com/forums/10711/entries/68916
2. Twitter blog: Making progress on spam. http://blog.twitter.com/2008/08/making-progress-on-spam.html (August 2008)
3. The 2000 following limit on Twitter. http://twittnotes.com/2009/03/2000-following-limit-on-twitter.html (Mar 2009)
4. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (Oct 1999)
5. Bonato, A., Janssen, J., Pralat, P.: A geometric model for on-line social networks. In: WOSN (Jun 2010)
6. Catone, J.: Twitter's follow limit makes Twitter less useful. http://www.sitepoint.com/blogs/2008/08/13/twitter-follow-limit-makes-twitter-less-useful/ (Aug 2008)
7. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Review 51(4), 661–703 (2009)
8. Ghosh, S., Korlam, G., Ganguly, N.: The effects of restrictions on number of connections in OSNs: A case-study on Twitter. In: Workshop on Online Social Networks (WOSN) (Jun 2010)
9. Guclu, H., Yuksel, M.: Scale-free overlay topologies with hard cutoffs for unstructured peer-to-peer networks. In: IEEE ICDCS (2007)
10. Java, A., Song, X., Finin, T., Tseng, B.: Why we Twitter: understanding microblogging usage and communities. In: WebKDD / SNA-KDD. pp. 56–65 (2007)
11. Krapivsky, P.L., Rodgers, G.J., Redner, S.: Degree distributions of growing networks. Phys. Rev. Lett. 86(23), 5401–5404 (Jun 2001)
12. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about Twitter. In: WOSN (2008)
13. Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. In: ACM KDD. pp. 611–617 (2006)
14. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: ACM WWW. pp. 591–600 (2010)
15. Mislove, A., Koppula, H.S., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Growth of the Flickr social network. In: WOSN (2008)
16. Mitra, B., Dubey, A., Ghose, S., Ganguly, N.: How do superpeer networks emerge? In: IEEE INFOCOM. pp. 1514–1522 (2010)
17. Owyang, J.: The many challenges of social network sites. http://www.web-strategist.com/blog/2008/02/11/the-many-challenges-of-social-networks/ (Feb 2008)