ESTIMATING NETWORK DEGREE DISTRIBUTIONS UNDER SAMPLING: AN INVERSE PROBLEM, WITH APPLICATIONS TO MONITORING SOCIAL MEDIA NETWORKS

By Yaonan Zhang[†], Eric D. Kolaczyk[†] and Bruce D. Spencer[‡],

Boston University[†] and Northwestern University[‡]

Networks are a popular tool for representing elements in a system and their interconnectedness. Many observed networks can be viewed as only samples of some true underlying network. Such is frequently the case, for example, in the monitoring and study of massive, online social networks. We study the problem of how to estimate the degree distribution – an object of fundamental interest – of a true underlying network from its sampled network. In particular, we show that this problem can be formulated as an inverse problem. Playing a key role in this formulation is a matrix relating the expectation of our sampled degree distribution to the true underlying degree distribution. Under many network sampling designs, this matrix can be defined entirely in terms of the design and is found to be ill-conditioned. As a result, our inverse problem frequently is ill-posed. Accordingly, we offer a constrained, penalized weighted least-squares approach to solving this problem. A Monte Carlo variant of Stein's unbiased risk estimation (SURE) is used to select the penalization parameter. We explore the behavior of our resulting estimator of network degree distribution in simulation, using a variety of combinations of network models and sampling regimes. In addition, we demonstrate the ability of our method to accurately reconstruct the degree distributions of various sub-communities within online social networks corresponding to Friendster, Orkut, and LiveJournal. Overall, our results show that the true degree distributions from both homogeneous and inhomogeneous networks can be recovered with substantially greater accuracy than reflected in the empirical degree distribution resulting from the original sampling.

1. Introduction.

Many networks observed or investigated today are samples of much larger networks (Kolaczyk, 2009, Ch 5). Let $G = (V; E)$ be a graph representing a network, with vertex set V and edge set E. Similarly, let $G = (V; E)$ denote a subgraph of G, representing a part of the network obtained through some sort of network sampling. Although practitioners typically speak of the network when presenting empirical results, frequently

Partially supported by AFOSR award 12RSL042 and NSF grant CNS-0905565. This work was begun during the 2010–2011 Program on Complex Networks at SAMSI.

it is only a sampled version G (or some function there of, such as when sampling yields estimates of vertex degrees directly) of some true underlying network G that is available to them, either by default or design. A central statistical question in such studies, therefore, is how much the properties of the sampled network reflect those of the true network.

Sampling is of particular interest in the context of online social networks. One reason for such interest is that these networks are usually very large. For example, social networks from Friendster, LiveJournal, Orkut, and Amazon have been studied in Yang and Leskovec (2012) having, respectively, 117:7M; 4:0M; 3:0M and 0:33M vertices and 25861M , 34:9M , 117:2M and 0:92M edges. Similarly in Ribeiro and Towsley (2010), networks from Flickr and Youtube were studied having millions of vertices and edges as well. The large size of these social networks makes it costly querying the entire network, particularly if the goal is to monitor these networks regularly over time. In addition, the decentralized nature of many such networks frequently means that few { if any { people or organizations have complete access to the data.

The topic of network sampling goes back at least to the seminal work of Ove Frank and his colleagues, starting in the late 1960s and extending into the mid-1980s. See Frank (2005), for example, for a relatively recent survey of that literature. With the modern explosion of interest in complex networks, there was a resurgence of interest in sampling. Initially, the focus was on the simple awareness, and then understanding, of whether and how sampling affects the extent to which the shape of the degree distribution of the observed network G reflects that of the true network G. Seminal work during this period includes an important empirical study by Lakhina et al. (2003), in the context of traceroute sampling in the Internet, with followup theoretical work by Achlioptas et al. (2005), and work by Stumpf and colleagues (Stumpf and Wiuf, 2005; Stumpf, Wiuf and May, 2005, e.g.), motivated, among other things, by networks arising in computational biology.

The focus on sampling of online social networks, as described above, is arguably the most recent direction intrature.e, is examples in this highly active area include Ahn et al. (2007), Ahmed et al. (2010), Ahmed, Neville and Kompella (2011), Ahmed, Neville and Kompella (2012), Maiya and Berger-Wolf (2010a), Maiya and Berger-Wolf (2010b), Li and Yeh (2011), Yoon et al. (2011), Shi et al. (2008), Mislove et al. (2007), Lu and Bressan (2012), Lim et al. (2011), Gjoka et al. (2010), Gjoka et al. (2011), Wang et al. (2011), Zhou et al. (2011), Kurant et al. (2011), Kurant, Markopoulou and Thiran (2011), Salehi et al. (2011), Mohaisen et al. (2012), Jin et al. (2011).

In all of these papers, there is a keen interest in understanding the extent to which characteristics of the network G are reflective of those of G. Typical characteristics of interest include degree distribution, density, diameter, the distribution of the clustering coefficient, the distribution of sizes of weakly (strongly) connected components, Hop-plot, distribution of singular values (vectors) of the network adjacency matrix, the graphlet distribution, the vertex (edge) label density, and, the assortative mixing coefficient.

Here, in this paper, the network property we focus on is degree distribution. The degree distribution of a network G, denoted by $f f_d g$, specifies the proportion f_d of vertices to have exactly d incident edges, for $d = 0; 1;$. It ill-conditioned. As a result, the estimation of f must be handled with care, since naive inversion of ill-conditioned operators in inverse problems typically will inflate the `noise' accompanying the process of obtaining measurements, often with devastating effects on our ability to recover the underlying object (e.g., function or density). Here we offer, to the best of our knowledge, the first principled estimator of a true degree distribution f from a sampled degree distribution f . In particular, we propose a constrained, penalized weighted least squares estimator, which, in particular, produces estimates that are non-negative (by constraint) and invert the matrix $\tilde{P}$ in a stable fashion (by construction), in a manner that encourages smooth solutions (through a penalty).

The rest of the paper is organized as follows. In Section 2 we provide a detailed characterization of our inverse problem, discussing the nature of the operator and the distribution of noise. In Section 3 we describe our proposed approach to solving this inverse problem, including a method for the automatic selection of the penalization parameter. In Section 4 we provide results of a simulation study, in which we study the impact on the performance of our estimator of various parameters, including the total number of vertices, the density of the network, sampling rates and network types. In Section 5, we return to the primary application of interest here, that of monitoring online social networks. There we demonstrate the ability of our method to simultaneously reconstruct accurately the degree distributions of various sub-communities within online social networks corresponding to Friendster, Orkut, and LiveJournal. Finally, some additional discussion and conclusions may be found in Section 6.

2. Characterizing the Inverse Problem.

In solving inverse problems generally, it is important to understand the nature of both the operator and the noise. Here the operator, in the form of the matrix $\tilde{P}$, will derive entirely from the network sampling design. At the same time, the `noise' (or, more formally, the randomness in our measurements) also derives from the sampling design. This linking of both operator and noise to our sampling lends a certain element of uniqueness to our particular inverse problem, the nature of which we aim to characterize in this section.

2.1. Nature of the problem.

To begin with, assume we know the total number of vertices n_v in the underlying network. This is a reasonable assumption in the cases of, for example, sampling a phone call network, or surveying among a class of students for their interactions. It is also not unreasonable in the context of many online social networks where, for example, this may either be readily available to those who own the network or reported to the community as a basic summary statistic (e.g., the number of members with active pages on Facebook). Thus we know the degree distribution f if and only if we know the degree counts $N = (N_0; N_1; \quad ; N_M)$, where N_k

(i.e., high-frequency) at larger values of i. Since most degree distributions encountered in practice, as well those induced through common choices of random graph models (some examples of which we use in Section 4), are relatively smooth, typically with either exponential or power-law behavior in the tails, intuitively it is the first handful of right singular vectors upon which a sensible estimator should be based. The stability of this estimator can be summarized through the condition number of P

matrix. We consider them ordered from simpler to more complex. We refer readers to Kolaczyk (2009, Ch 5) for additional background on network sampling and a more comprehensive list of sampling designs.

2.2.1. *Ego-centric and one-wave snowball sampling.* Ego-centric sampling (also called unlabeled star sampling) is a simple, non-adaptive (conventional) sampling design. As Handcock and Gile (2010) write that \[a] sampling design is conventional if it does not use information collected during the survey to direct subsequent sampling of individuals. . . [and] a sampling design [is] adaptive if it uses information collected during the survey to direct subsequent sampling, but the sampling design depends only on the observed data." Under ego-centric sampling, first a set of vertices is selected according to independent Bernoulli(p) trials at each vertex. Then all edges incident to the selected vertices are observed. In this case, the operator P is a diagonal matrix with the sampling rate p at each diagonal position, i.e.,

$$P_{ego}(i;j) = \begin{cases} p & \text{for } i = j = 0;1; \quad ;M \\ 0 & \text{for } i;j = 0; \quad ;M;\ i \neq j\ : \end{cases} \tag{2.4}$$

A natural eyF30 108ui

designs we consider with known and constant matrix P would be the logical point of departure for research on correcting the sampling bias of the degree distribution in more complex adaptive designs.

For a diagonal P matrix, the singular values are equal to the diagonal elements. Both the left and right singular vectors are the canonical set of basis vectors $f e_i g_{i=1}^{M+1}$, where e_i contains a 1 at the ith entry and 0 at all the other entries. Since $P_{ego} = I \quad p$, where I is the identity matrix, P_{ego} is not ill-conditioned at all. To estimate the degree count vector N we need only scale the observed degree count vector N by $1=p$. That is, the naive estimator is $\hat{N}_{naive} = N \; =p$.

In one-wave snowball sampling, the observed degree counts are biased, because in the second round of vertex selection, there is more chance to select the vertices that have more connections. The observed degree count vector therefore can be thought of as moving to the right of the true degree count vector. Hence, at a minimum, a good estimator should correct the observations by moving the distribution back to the left. How difficult this task may be is summarized by the condition number of P_{snow}, which is equal to

$$\frac{P_{snow}(M;M)}{P_{snow}(0;0)} = \frac{1 \quad (1 \quad p)^{M+1}}{1 \quad (1 \quad p)} = \frac{1 \quad (1 \quad p)^{M+1}}{p}; \tag{2.6}$$

and therefore depends on the relationship between the expected proportion p of vertices sampled initially and the maximum degree M. In the case where p is fixed, as M increases, the condition number is upper bounded by $\frac{1}{p}$. On the other hand, if $M p = o(1)$, using the approximation $(1 \quad p)^{M+1} \quad 1 \quad (M+1)\, p$, we find that the condition number behaves as $(M + 1)$.

These observations suggest that, for instance, under low sampling rates the inverse problem is increasingly ill-posed for estimating degree distributions of heavier tails. Also, the bounds on the condition numbers suggest that, in contrast to estimation of the mean from a sample from a finite population, where the accuracy depends on the sample size rather than the fraction of the population that is sampled, for estimation of complex properties of networks the accuracy depends strongly on the fraction of the population that is sampled.

2.2.2. Induced and incident subgraph sampling. These two sampling designs are both non-adaptive and analogous in spirit, differing only in the order of selection of vertices and edges. In induced subgraph sampling, a set of vertices is selected as independent Bernoulli(p) trials (other variations are possible { see below). Then, all edges between selected vertices are observed, i.e., we observe the subgraph induced by this vertex subset. This sampling scheme has been used in the analysis of technological and biological networks (Stumpf and Wiuf, 2005). Conversely, under incident subgraph sampling we selectedgesas independent Bernoulli(p) trials and we then observe all vertices incident to at least one selected edge.

The P matrix for induced subgraph sampling is

$$P_{ind}(i;j) = \begin{cases} {}^{j}_{i}\, p^{i+1}(1 \quad p)^{j \quad i} & \text{for } 0 \quad i \quad j \quad M \\ 0 & \text{for } 0 \quad j < i \quad M \end{cases}; \tag{2.7}$$

while that for incident subgraph sampling is

$$P_{inc}(i;j) = \begin{cases} {}^{j}_{i}\, p^{i}(1 \quad p)^{j \quad i} & \text{for } 1 \quad i \quad j \quad M \\ 0 & \text{for } 0 \quad j < i \quad M \end{cases}: \tag{2.8}$$

Notice th[-466e-474(th[-4663(0)]4eomTJ/F1 /f 3.38a6e-4744[( )1 Td [( )]TJ/F30 10.9091 Ss(19m10

Fig 2. Singular values decay under Induced Subgraph sampling$M = 20$.

the expected behavior of this estimator. As can be seen from the illustration in Figure 3, the right singular vectors behave like a Fourier basis, in that

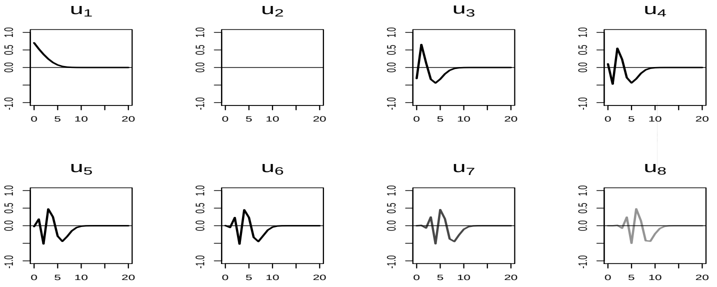

Fig 4 . The first 12 left singular vectors under Induced Subgraph sampling, ordered by singular values from big to small: maximum degree $M = 20$, sampling rate $p = 20\%$.

For one-wave snowball sampling, the representation (2.10) still applies. However, the indicator functions are not independent. Straightforward arguments yield that the covariance and variance of N_k for $k = 0; 1; \ldots; M$ are

$$\begin{aligned}\mathrm{Cov}(N_k; N_l) = &\sum_t N_{1klt}(1-p)^{k+l-t}\\ &+ \sum_t N_{0klt}(1-p)^{k+l-t+2} - N_k N_l (1-p)^{k+l+2};\end{aligned} \tag{2.11}$$

and

$$\begin{aligned}\mathrm{Var}(N_k) = &N_k P_{snow}(k; k)\\ &+ \sum_t N_{1kkt}(1-p)^{2k-t} + \sum_t N_{0kkt}(1-p)^{2k-t+2}\\ &+ N_k^2(1-p)^{2k+2} - N_k\left[1-2(1-p)^{k+1}\right];\end{aligned} \tag{2.12}$$

where N_{0klt} (N_{1klt}) is determined by the underlying network G, defined as the number of ordered pairs of nonadjacent (adjacent) distinct vertices of degrees k and l, respectively, which have t common adjacent vertices.

Now consider the marginal distributions of the N_k under snowball sampling and induced subgraph sampling. Note that the first term in (2.12) and (2.14) is the k-th entry of the expectation PN. This observation suggests that, if the remaining terms in the variance (as well as the off-diagonal terms corresponding to covariances) are sufficiently small, a Poisson model might again be acceptable.

More precisely, if the sampling rate p is small, then each of the indicators in (2.10) and (2.13) likely has only very small probability of being equal to one. On the other hand, if the graph is large (i.e., n_v is large) and k is not too far out in the tail of the distribution (i.e., k is not too close to M), then there should be many such indicators. So a Poisson approximation would make sense here. Given, however, that these indicator variables are dependent, the necessary argument is somewhat more involved. We present a formal justification, using the Chen-Stein method, in Appendix B.

Simulation can be used to assess this approximation. Some representative results, shown in Figure 5, confirm the reasonableness of a Poisson approximation for the marginal distribution of the N_k, under induced subgraph sampling, for k within a reasonable distance from the mean.

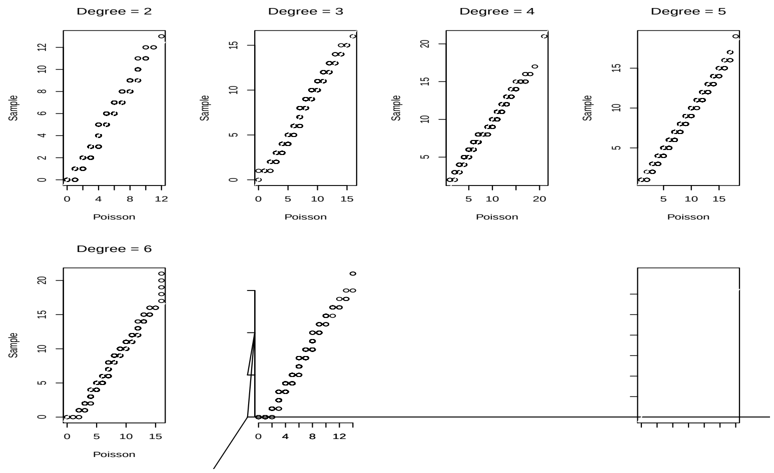

Fig 5. QQ plot: distribution of N_i compared to Poisson distribution with mean $(PN)_i$. The underlying network is ER with $n_v = |V| = 1000$ and $n_e = |E| = 50000$. Sampling rate $p = 5\%$. The average degree of sample is equal to 5.

In summary, for all of the sampling plans considered in this paper, an approximate Poisson marginal distribution is arguably reasonable for the observed counts N_k. Thus, a Poisson regression model is suggested for solv-

that is, f_k f_l if k and l are close. Examples of networks with smooth degree distributions include Erdos-Renyi (ER), mixture of ER, power-law

We define a weighted mean square error (WMSE) in the observation space as

$$\mathrm{WMSE}(\hat{N};N) = E\left[(PN \quad P\hat{N})^T C \;{}^{1}(PN \quad P\hat{N})\right] : \tag{3.3}$$

Under the conditions that $f\ \ (N\ \ )$ is weakly differentiable and that $E\,|f\ \ (N\ \ )|$ is bounded (which we verify following the arguments in Appendix C), a generalized SURE estimate for the WMSE can be obtained as

$$\begin{aligned}\widehat{\mathrm{WMSE}}(\hat{N};N) = (\ PN)^T C \;{}^{1}PN + (\ P\hat{N})^T C \;{}^{1}P\hat{N} \\ +\,2 \quad \left(\mathrm{Trace} \quad P\frac{\partial\hat{N}}{\partial N} \right) \\ 2(P\hat{N})^T C \;{}^{1}N \quad : \end{aligned} \tag{3.4}$$

The first term in (3.4) involves the unknown N. However, we may drop this term because it does not involve . The last three terms have $\hat{N}$ in them, which is a function of

3.3. Approximation of the covariance matrix C. For the ego-centric sampling design, recall that the N_k are independent random variables, distributed according to a binomial with parameters p and N_k. As a result, the covariance matrix C is simply $p(1 \quad p) \quad \mathrm{diag}(N)$. In contrast, for the one-wave snowball sampling and the induced subgraph sampling (as well as the related incident subgraph and random walk sampling), C will have non-zero off-diagonal elements. Recall, however, that these off-diagonal elements involved higher-order properties of the graph, in the sense of summarizing even more structure than the degree distribution we seek to estimate. Accordingly, it is unrealistic to think to incorporate this information into our estimation strategy. We instead focus on the diagonal elements of C.

We approximate the covariance matrix C with a diagonal matrix of the form

$$\hat{C} = \mathrm{diag}(N_{smooth}) + \tag{3.6}$$

4.1. Design. There are several parameters that need to be chosen with some care. Here we list them and discuss the conventions we applied.

b: The random vector b must have zero mean, covariance matrix I , and bounded higher order moments; here we use a multivariate normal, i.e. b $N(0; I)$.

: In principle, the value should be small enough to approximate of networks are studied: those from the Erd•os-Renyi model and those from a block model with two blocks. These are two basic models commonly used in network studies (e.g., Kolaczyk 2009, Ch 6). In the Erd•os-Renyi model, edges are randomly assigned to each pair of vertices with a given probability, sampling (Figure 9), the block model has a broader range of degrees than the Erdős-Rényi model at any given choice of our other simulation parameters. In (2.13), for each k, the indicator function involving $u \in V$ with higher d_u has lower probability of being equal to 1. Thus a better Poisson approximation of N_k and a more accurate approximation of C occur under the block model. A power-law network has an even broader degree distribution. For the same reasons, therefore, we expect the estimators for the power-law like networks in the applications of Section 5 to perform similarly well. However, the results for Erdős-Rényi and the block model are quite close in Figure 7 and Figure 8. This is because only the vertex with degree k in the true network can possibly contribute to degree k under ego-centric and one-wave snowball sampling.

Three sampling rates are studied: 10%, 20%, and 30%. Our results show that there is less accuracy for smaller sampling rate, as is to be expected. In the literature on Internet community monitoring, 30% sampling rates have been suggested as reasonable for preserving network properties to a reasonable accuracy (Leskovec and Faloutsos, 2006). In our results, we see that our estimators of degree distribution perform fairly well based on as low as a 10% sampling rate.

Fig 7 . Simulation results for ego-centric sampling. Error measured by K-S D-Statistic. For each sampling rate, the three boxes from left to right represent K-S D-Statistic comparing the true degree distribution with (left) sample degree distribution, (middle) estimated degree distribution using the non-parametric method, and (right) estimated degree distribution using the proposed method. (Online versions of figure are in color.)

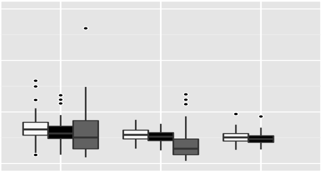

Fig 8 . Simulation results for one-wave snowball sampling. Error measured by K-S D-Statistic. For each sampling rate, the three boxes from left to right represent K-S D-Statistic comparing the true degree distribution with (left) sample degree distribution, (middle) estimated degree distribution using the non-parametric method, and (right) estimated degree distribution using the proposed method. (Online versions of figure are in color.)

5. Applications. The cost of any sampling strategy varies with the structure of the network and the protocol. As we have remarked, sampling is of particular interest in the context of online social networks. In online social networks where each user is assigned an unique user id, it is a common practice to select a set of users by querying a set of randomly generated user id's (Ribeiro and Towsley, 2010). Thus our induced subgraph sampling can be applied there. In this section, we use our degree distribution estimation method on data from three online social networks: Friendster, Orkut, and LiveJournal. These data are available on the SNAP (Stanford Network Analysis Project) website. In the following we present our estimates of various degree distributions from these online social networks. In addition, we show how these degree distributions help us to gain insight about the epidemic thresholds of these networks, which is relevant to the concept of social influence, spread of rumors and viral marketing.

5.1. Estimateing degree distributions from online social networks. It is now well-understood that large-scale, real-world networks frequently have heavy-tailed degree distributions. Stumpf and Wiuf (2005) proved analytically that for a network with an exact power-law degree distribution, although its sampled network under our sampling method (induced Subgraph

ing, life/style, life/support, sports, student life and technology" (Yang and Leskovec, 2012). It is the degree distributions for subnetworks corresponding to collections of ground-truth communities such as these that we estimate here.

Figure 10 gives an example of the estimators. The first row is for three sub-networks from Friendster. Communities are ordered according to the number of users in them. In the top-left subplot, vertices from the top 5 communities form an induced sub-network for which the degree distribution is to be estimated. Then Bernoulli sampling of vertices with 30% sampling rate is performed on this sub-network, and our estimation method is applied. Similarly, the true network in the top-middle plot is induced by top 6-15 communities, and in the top-right plot the true network is induced by the top 16-30 communities. The second row and the third row show estimates of Orkut and LiveJournal respectively. Examination of these plots shows that, while the sampled degree distribution can be quite off from the truth, particularly in the case of the Friendster and Orkut networks, correction for sampling using our proposed methodology results in estimates that are nearly indistinguishable by eye from the true degree distributions.

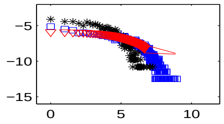

Fig 10 . Estimating degree distributions of communities from Friendster, Orkut and Live-journal. Squares represent the true degree distributions, dots represent the sample degree distributions, triangles represent the estimated degree distributions. Sampling rate=30%. Points which correspond to a density $< 10^{-4}$ are eliminated from the plot. (Online versions of figure are in color.)

In Table 1, the median and inter-quartile range are computed based on the application of our estimator to 20 samples. The estimated degree distribution greatly improves over the degree distribution of the sample, as measured by K-S D-statistic. In fact, the improvement in accuracy is by an order of magnitude, with the values of the D-statistic produced by our estimator being on the same order of magnitude as the best results in our simulation study.

degree, M_2 be the second raw moment of the degree distribution, $n_e = jEj$ be the number of total edges, and $U = (2\quad n_e(n_v\quad 1)=n_v)^{1=2}$. Then we have the following relationship,

$$M_1 \qquad \sqrt{M_2} \qquad {}_1 \qquad U\ : \tag{5.1}$$

The proof of the first two inequalities can be found in Van Mieghem (2011), and the third (upper bound) can be found in Lovasz (1993). Thus we have the bounds for the epidemic threshold $_c$,

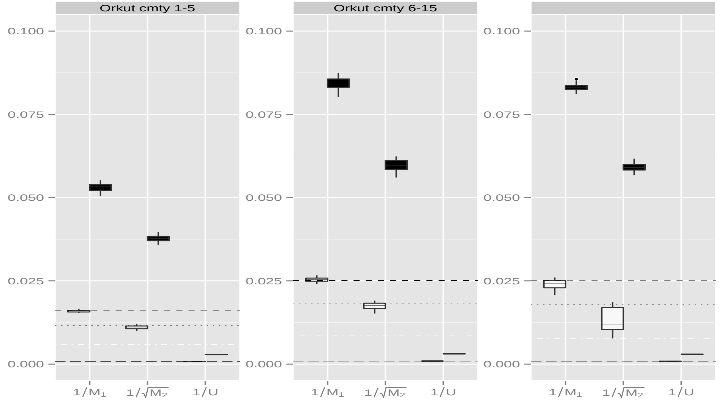

Fig 12 . Bounds for the epidemic spreads of Orkut networks, each box is estimated based on 20 samples, four horizontal lines are the true values for $\frac{1}{M_1}$, p$\frac{1}{M_2}$, $\frac{1}{1}$ and $\frac{1}{U}$ from top to bottom. For each bound, the two boxes from left to right correspond to the estimated value using (left) the proposed method and (right) the sample degree distribution. (Online versions of figure are in color.)

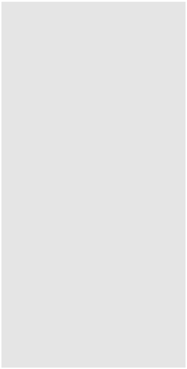

Fig 13 . Bounds for the epidemic spreads of LiveJournal networks, each box is estimated based on 20 samples, four horizontal lines are the true values for$\frac{1}{M_1}$, $\frac{1}{}$

2.1

radius) of the network can be successfully bounded by functions of our es-

The equation at the (k 1)th row is

Finally,

$$x_{k \quad n} = (\quad 1)^n \binom{k \quad 1}{n} \tag{A.11}$$

Therefore, the entries in the kth eigenvector are

$$\text{ʉ}_k(j) = \begin{cases} (\quad 1)^{k \quad j} \binom{k \quad 1}{j \quad 1} & \text{for } 1 \quad j \quad k \\ 0 & \text{fork} < j \quad M + 1 \end{cases} \tag{A.12}$$

The theorem is true for k by k matrix P.

APPENDIX B: POISSON APPROXIMATION

Here we give a proof of the Poisson approximation of the cumulative de-

Proof of Theorem B.1 We sketch the proof briefly here. Without loss of generality, (partially) order the vertices $f v_1; : : : ; v_{n_v} g$ by (non)decreasing degree. Associate a binary random vector $X(_1; : : : ; X_{n_v})$ with the vertices, where the elements are independent Bernoulli random variables with parameter p. So X represents the selection of vertices for inclusion in S in the case of induced subgraph sampling and the initial selection of vertices in the case of snowball sampling. Now let $I_{v;k}$ be an indicator random variable, which is one if v *2* S and d_v k. Then the variables $I_{v;k}$ are so-called `increasing functions' of realizations of X. So Corollary 2.E.1, page 28, of Poisson Approximation, by Barbour and colleagues, yields our result.

In more detail, there are two key observations to be made. First, we need the $I_{v;k}$ to be increasing functions. This induces positive correlation among these indicator variables and it makes a general Chen-Stein bound become much cleaner, as in our theorem, in that it can be expressed explicitly in terms of means and variances. Partial ordering means that if we let x and y be two possible realizations of X, then x y if and only if x_i y_i for all i. And a function f is increasing if $f(x)$ $f(y)$ whenever x y. For x to be less than or equal to y, it suffices to think of what happens simply when a new vertex enters the sample S. One element of x will change from a zero to a one, so x y. What happens to $I_{v;k}$? If v is a vertex that was already in S, under x, then adding a vertex to the sample under y can either not change or increase its degree. So $I_{v;k}(x)$ $I_{v;k}(y)$. On the other hand, if v itself was the new vertex to enter S under y, the same statement can be made.

Second is the observation that elements of X are independent in our setting, which is guaranteed by our assumption of Bernoulli sampling. Taken

Ahmed, N. K. , Neville, J. and Kompella, R. R. (2012). Network Sampling Designs for Relational Classification. In ICWSM .

Ahmed, N. K. , Berchmans, F. , Neville, J. and Kompella, R. (2010). Time-based sampling of social network activity graphs. In Proceedings of the Eighth Workshop on Mining and Learning with Graphs 1{9. ACM.

Ahn, Y.-Y. , Han, S. , Kwak, H. , Moon, S. and Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. In Proceedings of the 16th international conference on World Wide Web 835{844. ACM.

Bailey, N. T. et al. (1975). The mathematical theory of infectious diseases and its applications. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Boyd, S. and Vandenberghe, L. (2004). Convex optimization. Cambridge university press.

Cochran, W. (1977). G.(1977); Sampling techniques. New York, Wiley and Sons 98 259{261.

CVX Research, I. (2012). CVX: Matlab Software for Disciplined Convex Programming, version 2.0 beta. http://cvxr.com/cvx .

Daley, D. and Gani, J. (1999). Epidemic modelling: An introduction.

Dong, J. and Simonoff, J. S.

Computer Society Symposium on 343{359. IEEE.

Kolaczyk, E. D. (2009). Statistical analysis of network data. Springer.

Kurant, M. , Markopoulou, A. and Thiran, P. (2011). Towards unbiased BFS sampling. Selected Areas in Communications, IEEE Journal on 29 1799{1809.

Kurant, M. , Gjoka, M. , Butts, C. T. and Markopoulou, A. (2011). Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems

Salehi, M. , Rabiee, H. R. , Nabavi, N. and Pooya, S. (2011). Characterizing Twitter with Respondent-Driven Sampling. In Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on 1211{1217. IEEE.

Shi, X. , Bonner, M. , Adamic, L. A. and Gilbert, A. C. (2008). The very small world of the well-connected. In Proceedings of the nineteenth ACM conference on Hypertext and hypermedia 61{70. ACM.

Stumpf, M. P. and Wiuf, C. (2005). Sampling properties of random graphs: the degree distribution. Physical Review E 72 036118.

Stumpf, M. P. , Wiuf, C. and May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. Proceedings of the National Academy of Sciences of the United States of America 102 4221{4224.

Van Mieghem, P. (2011). Graph spectra for complex networks. Cambridge University Press.

Van Mieghem, P. , Omic, J. and Kooij, R. (2009). Virus spread in networks. Networking, IEEE/ACM Transactions on 17 1{14.

Wang, T. , Chen, Y. , Zhang, Z. , Xu, T. , Jin, L. , Hui, P. , Deng, B. and Li, X. (2011). Understanding graph sampling algorithms for social network analysis. In Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on 123{128. IEEE.

Yang, J. and Leskovec, J. (2012). De ning and evaluating network communities based on ground-truth. In Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics

Kose 0 G mE5.3 076(S.-H3(vec,)-373(J.)]TJ/F32 8..9664 Tf 29.95 0 Td [(,)]TJ/F366.9664 Tf 17im.3 076(K.-)89(vi,)-3