An approach to correct biases induced by snowball sampling

Johannes Illenberger*, Gunnar Flötteröd†and Kai Nagel‡

August 1, 2008

Words: 3577
Figures: 6

*Johannes Illenberger (corresponding author) is with the group of Transport Systems Planning and Transport Telematics, Institute for Sea- and Land-Transport, Technical University Berlin, Germany, Salzufer 17-19, 10587 Berlin, +49 30 31478793, illenberger@vsp.tu-berlin.de

†Gunnar Flötteröd is with the group of Transport Systems Planning and Transport Telematics, Institute for Sea- and Land-Transport, Technical University Berlin, Germany, Salzufer 17-19, 10587 Berlin, +49 30 31429520, floetteroed@vsp.tu-berlin.de

‡Kai Nagel is the head of the group of Transport Systems Planning and Transport Telematics, Institute for Sea- and Land-Transport, Technical University Berlin, Germany, Salzufer 17-19, 10587 Berlin, +49 30 31423308, nagel@vsp.tu-berlin.de

Abstract

This article treats an issue that arises when statistical properties of a social network are estimated from data that has been obtained by snowball sampling. Snowball sampling tends to over-represent vertices with high degrees, which leads to a bias in the estimates of all statistical properties that correlate with the degree. We propose a simple method to compensate this bias. The method is tested by simulating a snowball sampling on a co-authorship network. The results show that our approach leads to reasonable estimates of the mean degree and the clustering coefficient.

1 Introduction

In recent years a significant amount of research in the field of complex networks has been conducted. The increasing availability of real-world network data motivates interdisciplinary research in several fields such as physics, mathematics, computer science and sociology. In sociology, the analysis of social networks has become an important matter and a number of tools for statistical analysis including Statnet [7], Pajek [2] or UCINET [1] have been developed.

Snowball sampling defines an iterative survey procedure where (i) an initial set of persons is more or less randomly selected and asked to report their social connections, and (ii) in every following iteration, the interviewed individuals are selected from the set of those people who have been reported as social contacts for the first time in the previous iteration. The structure of the social networks under investigation affects which people are interviewed, that is, the overall set of interviewed people is not a random sample from the entire population. This article describes techniques that avoid biases in the estimation of network parameters from suchlike conducted surveys.

A first analytical discussion of snowball sampling has been provided by Goodman [6]. That article discusses the impact of the two main parameters of the snowball sampling algorithm: the number of contacts named by an interviewed individual and the number of iterations conducted. Lee et. al [9] investigated in the statistical properties of sampled networks. Beside snowball sampling, node and link sampling were conducted on computer, biological and social networks. However, the impossibility to capture a complete network is only one of several issues related to snowball sampling. In the particular case of social networks, one also has to deal with non-response effects. A detailed sensitivity analysis of such effects is given by Kossinets [8].

The remainder of this article is organized as follows. Section 2 describes the snowball sampling procedure in greater detail. Section 3 presents the main results of this article. It introduces several network statistics that are to be estimated from snowball sampled data. For each statistic, it discusses the nature of the bias (possibly) introduced by snowball sampling, proposes a technique to avoid this bias, and demonstrates the efficiency of the technique for an example network. Finally, Section 4 summarizes the article, and gives an outlook on upcoming problems that have not been addressed in this article.

2 Sampling method

2.1 Definition

We model a social network as an undirected and unweighted graph where the vertices represent individuals and the edges represent relations among the individuals (such as friendship, collaboration or some other kind of interaction). Denote the set of vertices by $\mathscr{V}$, its size by N, and the set of vertices that are linked to $v \in \mathscr{V}$ by an edge by $\mathscr{K}(v)$. The individual represented by a vertex is also called *ego*, and the vertices that are linked to an ego are called its *neighbours* or *alters*.

Snowball sampling is an iterative survey technique that aims to reveal structural information about a network by purposefully sampling a subset of its vertices and edges. By sampling, we mean some kind of interview with the person who is represented by the according vertex. We begin our presentation with a formal specification in Algorithm 1 and an illustration in Fig. 1.

Algorithm 1 Snowball Sampling

1. Initialize iteration counter $i = 0$.
2. Select a random sample $\mathscr{V}^{(0)}$ of vertices from the network.
3. Repeat as often as desired:
   (a) Increase i by 1.
   (b) Ask every ego v in $\mathscr{V}^{(i-1)}$ to report its alters. Let $\mathscr{V}^{(i)}$ contain all reported alters which have not been identified before. That is,

$$\mathscr{V}^{(i)} = \bigcup_{v \in \mathscr{V}^{(i-1)}} \mathscr{K}(v) - \mathscr{V}(0) - \ldots - \mathscr{V}(i-1).$$

Snowball sampling requires two basic parameters: (i) the number of initial vertices selected in Step 2 of Alg.

1 and (ii) the number of alters named by an ego in Step 3b. What kind of alters are to be named is defined by the so called *name generator* which is a question posed to the respondents. For example, the name generator may ask the respondents to name friends they discussed important matters with in the last six month. If the name generator includes distant alters, a high-degree network is sampled, and if the name generator only includes close friends, a low-degree network can be expected. That is, the name generator affects the degree of the sampled network.

If the network has disconnected components, a full coverage is only possible if at least one seed vertex per component is selected.

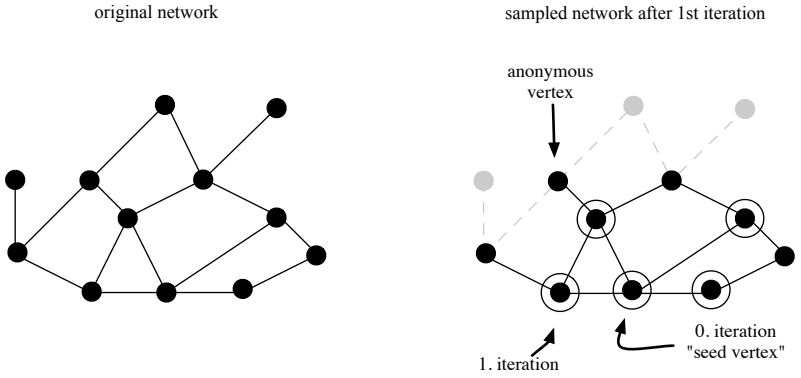

Figure 1: Snowball sampling mechanism. Left: original network, right: sampled network after the first iteration. Black and framed vertices are sampled, black vertices without frame are so called *anonymous* vertices. That means, vertices the existence of which is known, but which have not been sampled yet (where sampling means some kind of interview). Grey vertices are still uncovered.

2.2 Compensating the bias in snowball sampling

Networks sampled with snowball sampling tend to overestimate the mean degree in early iterations. By definition, high-degree vertices are more likely to be selected by this algorithm. In the 0th iteration, the set of sampled vertices, i.e., the seed vertices, is representative for the network if it is chosen at random. In the following iterations, vertices with higher degrees have a higher probability to be sampled compared to vertices with lower degrees because high-degree vertices have more connections along which they can be discovered. In addition, vertices that are predominantly connected to vertices with high degrees are also more likely to be selected. This effect is eminently distinct with networks that have right-skewed degree distributions like networks with power-law distributions [9].

Figure 2a visualizes this effect. It plots the fraction of sampled vertices with a particular degree over the sampling iteration. The higher the degree, the faster rises the according curve. Once all vertices are sampled, the over-representation of high-degree vertices ceases. However, the latter observation is of little practical use since the effort to completely sample a network is typically not affordable (and it also renders the snowball sampling technique itself superfluous).

It is possible to correct this kind of bias if we account for the probability with which a vertex of a particular degree is sampled. We cannot calculate this probability directly, but we can make a reasonable estimate in dependence of the vertex' degree and the iteration in which it is sampled. Let v be our vertex of interest and let k_v be its degree. The probability $P^{(i)}(v)$ that v is sampled before or in iteration i equals the probability that at least one of its neighbours is sampled before or in iteration $i-1$. This is the same as one minus the probability that none of its neighbours is sampled before or in iteration $i-1$, i.e., $P^{(i)}(v) = 1 - \prod_{w \in \mathcal{K}(v)} (1 - P^{(i-1)}(w))$ where we make the assumption that v's neighbours are discovered independently. Since $P^{(i-1)}(w)$ is just as unknown as $P^{(i)}(v)$, we approximate $P^{(i-1)}(w) \approx n^{(i-1)}/N$. Recall that N is the total number of vertices in the network and $n^{(i-1)}$ is the number of sampled vertices after iteration $i-1$. Assuming that the neighbours of v are discovered independently, we obtain $P^{(i)}(v) = 1 - (1 - n^{(i-1)}/N)^{k_v}$ as the probability that a particular vertex v of degree k_v is sampled before or in iteration i. Since this formula holds equally for all vertices of a particular degree, we subsequently denote by

$$P_k^{(i)} = 1 - \left(1 - \frac{n^{(i-1)}}{N}\right)^k \tag{1}$$

the probability that a vertex of degree k is discovered before or in iteration i. This expression is only dependent on the number of sampled vertices in the previous iteration $i-1$, and thus it allows to correct the estimated

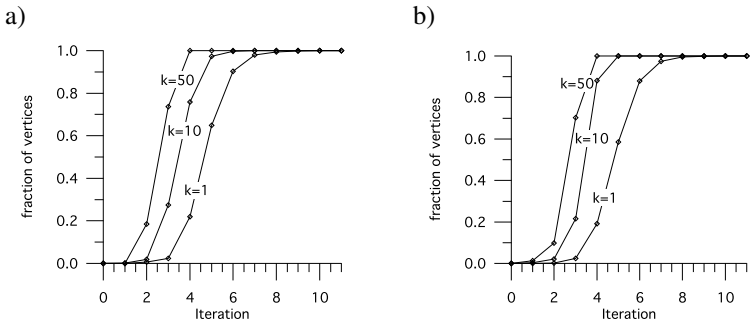

Figure 2: Fraction of vertices over sampling iteration for degree 1, 10 and 50. a) $n_k^{(i)}/N_k$ where $n_k^{(i)}$ is taken from simulation and N_k is known, b) calculated with Eq. 1, so that $n_k^{(i)}/N_k = P_k^{(i)} n^{(i)}/N$ where $n^{(i)}$ is taken from simulation. N_k denotes the total number of vertices with degree k, $n^{(i)}$ is the sampled number of vertices and $n_k^{(i)}$ denotes the sampled number of vertices with degree k.

degree distribution after every iteration i by weighting the number of sampled vertices of degree k by $1/P_k^{(i)}$ when estimating the frequency of their occurrence in the network.

3 Statistical properties of snowball-sampled networks

We investigate the characteristics of sampled networks by simulating a snowball sampling. After each iteration, we compute several statistical properties of the sampled network and compare them to the original network. Simulations are repeated 50 times with different random seeds, so that the set of ten seed vertices differs in each simulation run. We use the collaboration network of the condensed matter e-print archive (arxiv.org) [10]. For practical use, we extracted the giant component of the network including 36458 vertices and 171736 edges.

3.1 Mean degree and degree distribution

Figure 3 depicts the estimated mean degree $\langle k \rangle$ for the arxiv.org collaboration network. In a), it is calculated directly from the raw sampled data and in b), it is calculated using the bias correction of (1), that is,

$$\langle k \rangle^{(i)} \approx \frac{\sum_{v \in \mathscr{V}^{(0\ldots i)}} k_v / P_{k_v}^{(i)}}{\sum_{v \in \mathscr{V}^{(0\ldots i)}} 1/P_{k_v}^{(i)}} \tag{2}$$

where $\mathscr{V}^{(0\ldots i)} = \mathscr{V}^{(0)} \cup \ldots \cup \mathscr{V}^{(i)}$ denotes the set of all vertices sampled before or in iteration i, which is of size $n^{(0\ldots i)}$. The plot without bias correction (Fig. 3a) exhibits a considerably overestimation of the mean degree up to iteration four. If we refer to Fig. 2, we realize that up to iteration four predominantly vertices with degree greater than ten are sampled, whereas low-degree vertices are not sampled in a decisive amount yet. In contrast, in the plot with bias correction (Fig. 3b) the estimates' median value approximates the real mean degree very well already in iteration two.

Like many other real-world networks the arxiv.org co-authorship network follows a power-law degree distribution $P(k) \sim k^{-\gamma}$. The degree exponent γ can be extracted using the maximum likelihood estimator [3]

$$\gamma^{(i)} = 1 + \frac{1}{\frac{1}{n'^{(0\ldots i)}} \sum\limits_{v \in \mathscr{V}'^{(0\ldots i)}} \ln \frac{k_v}{k_{min}}} \tag{3}$$

where $\mathscr{V}'^{(0\ldots i)}$ denotes the set of all vertices sampled before or in iteration i that follow the power-law distribution, which is of size $n'^{(0\ldots i)}$ and k_{min} is the smallest degree for which the power-law holds. The evolution of $\gamma^{(i)}$ is shown in Fig. 3c. Due to the bias of the snowball sampling the tail of the power-law distribution develops much faster when compared to the remaining part of the distribution. Consequently, the γ exponent is underestimated in the first iterations. We can make use of Eq. 1 to take the sampling probability into account and modify the γ

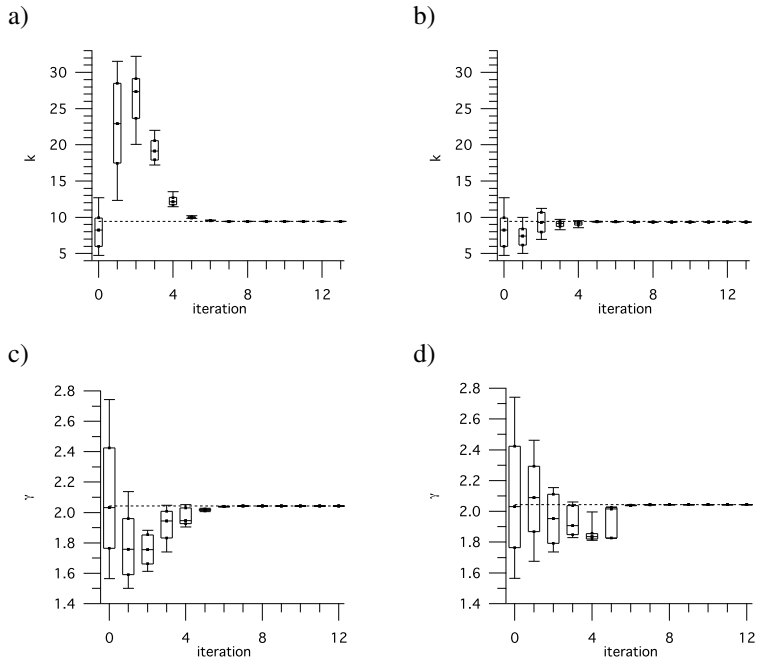

Figure 3: Mean degree: a) as measured from the sampled network, b) obtained with Eq. 2. Power-law exponent: c) obtained with Eq. 3 and d) with Eq. 4.

estimator to

$$\gamma^{(i)} = 1 + \frac{1}{\frac{1}{\sum_{v \in \mathscr{V}'(0\ldots i)} 1/P_{k_v}^{(i)}} \sum_{v \in \mathscr{V}'(0\ldots i)} \left(\ln \frac{k_v}{k_{min}} \right) / P_{k_v}^{(i)}}. \tag{4}$$

With this estimator, the decrease of the degree exponent is not that pronounced (Fig. 3d), however, the estimations for iteration four and five are much better with Eq. 3 (Fig. 3c) compared to Eq. 4.

3.2 Clustering, mutuality and degree correlation

For the clustering coefficient C of a network we use the average of the vertices' local clustering coefficients C_v. The clustering coefficient of a vertex is defined as

$$C_v = \frac{2y_v}{k_v\,(k_v - 1)} \tag{5}$$

where y_v denotes the number of edges between v's neighbours. In words, the clustering coefficient can be seen as the probability that a friend of one's friend is also one's friend, or the probability that a connected triple centered at v is closed to a triangle.

At this point we introduce the notion of *anonymous* vertices. Consider an outgoing edge of a sampled vertex v_1. We know that at the opposing end of this edge another vertex v_2 must exist even if v_2 has not been sampled yet. Imagine the following situation: Vertices A and B are sampled. A names C, D and F as its alters and B names F, G and H as its alters. Even if vertex F has not been sampled yet, we know that A and B have an alter in common and that there is at least one path from A to B. We denote F as "anonymous" since we know of the existence of F without having sampled it. Anonymous vertices can help to provide a better estimate of statistical properties of the sampled network. In the case of the clustering coefficient, the quantity y may be better approximated by including anonymous vertices. However, note that when averaging over the compete population anonymous vertices are excluded since they do not count as sampled vertices. In that manner $C^{(i)} = \frac{1}{n^{(0\ldots i)}} \sum_{v \in \mathscr{V}(0\ldots i)} C_v$.

From a theoretical point of view, it should be possible to determine the clustering coefficient within two iterations since, within two iterations, the snowball sampling is able to detect if a connected triple is closed to a triangle or not. The ratio of closed and open triples should provided a reasonable estimate of the clustering coefficient. However, in Fig. 4a it can be observed that the clustering coefficient is underestimated for the first iterations and about four iterations are required to approximate the real value.

Many real-world networks exhibit a correlation between clustering and degree, which is shown in Fig. 4c for the arxiv.org network. This correlation causes the bias of the snowball sampling to also affect the estimation of the clustering coefficient since high-degree vertices with potentially low clustering coefficients are preferably selected. Consequently, we apply the same probability weighting mechanism as for the mean degree to the clustering coefficient:

$$C^{(i)} = \frac{\sum_{v \in \mathscr{V}(0\ldots i)} C_v / P_{k_v}^{(i)}}{\sum_{v \in \mathscr{V}(0\ldots i)} 1 / P_{k_v}^{(i)}}. \tag{6}$$

Figure 4b depicts the clustering coefficient obtained with Eq. 6. It shows a much better approximation to the real clustering coefficient. As expected above, the values in the first iteration are already are reasonable estimates. A calculation of the clustering coefficient in the 0th iteration is meaningless.

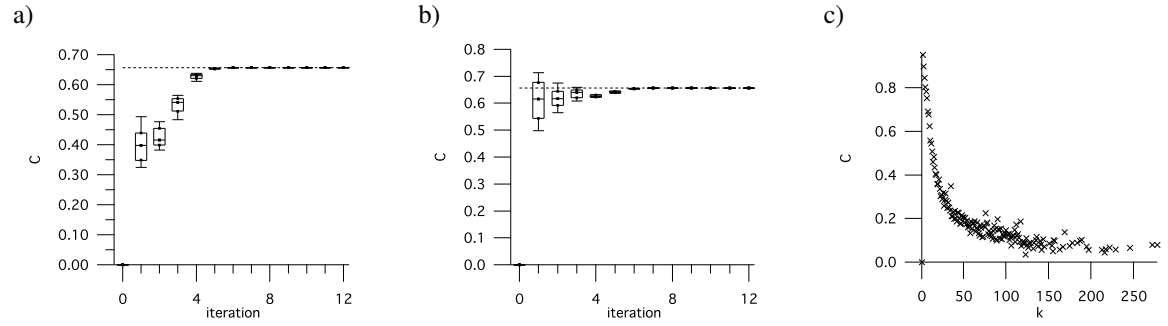

Figure 4: Clustering coefficient: a) as measured from the sampled networks, b) obtained with Eq. 6 and c) clustering-degree-correlation.

Beside triangles, relationships that form squares are also important structural properties. The density of squares in a network is measured by the quantity called mutuality which is defined as

$$\mu^{(i)} = \frac{\sum_{v \in \mathscr{V}(0\ldots i)} n_{2,v}}{\sum_{v \in \mathscr{V}(0\ldots i)} \sigma_{2,v}(w)} \tag{7}$$

where $n_{2,v}$ denotes the number of second neighbours of v, i.e., vertices that are two steps away form v, and $\sigma_{2,v}(w)$ denotes the number of paths of length two to v's second neighbours denoted with w. If $M = 1$, there exists only one path to a vertex's second neighbour (e.g. in a tree-like network). As M decreases, the amount of squares in the network increases, i.e., there exists more than one path to a vertex's second neighbour.

We can include probability weighting also for mutuality which leads to

$$\mu^{(i)} = \frac{\sum_{v \in \mathscr{V}(0\ldots i)} n_{2,v} / P_{k_v}^{(i)}}{\sum_{v \in \mathscr{V}(0\ldots i)} \sigma_{2,v}(w) / P_{k_v}^{(i)}}. \tag{8}$$

In principle, one can expect to need two iterations for a reasonable estimate of the mutuality. However, since the snowball sampling has been initiated with multiple seed vertices it is possible that some ego-centric networks already connect in the first iteration. In such cases a computation of the mutuality is already possible in the first iteration. Figure 5a depicts the mutuality values obtained by Eq. 7. The results scatter around the real value and a good estimate can be made from iteration four on. If we use Eq. 8 the estimation slightly improves (Fig. 5b) and iteration two already provides a reasonable median value.

Degree correlation, also known as assortativity, is defined as the Pearson correlation coefficient of the degrees of two vertices connected by a common edge. For a given network, the degree correlation can be calculated with the formula [11]

$$r = \frac{M^{-1} \sum_m j_m k_m - \left(M^{-1} \sum_m \frac{1}{2}(j_m + k_m)\right)^2}{M^{-1} \sum_m \frac{1}{2}(j_m^2 + k_m^2) - \left(M^{-1} \sum_m \frac{1}{2}(j_m + k_m)\right)^2} \tag{9}$$

where j_i, k_i are the degrees of the vertices at the ends of edge m and M is the total number of edges. The value of r lies in the range [-1,1], where positive values mean that vertices with high degrees tend to connect to other vertices with high degrees. Positive degree correlations, also called assortative mixing, is observed in many social networks including the arxive.org collaboration network.

Figure 5b depicts the degree correlation. Overall, the sampled networks show to be less assortative than the original network. Reasonable estimates cannot be made until iteration five.

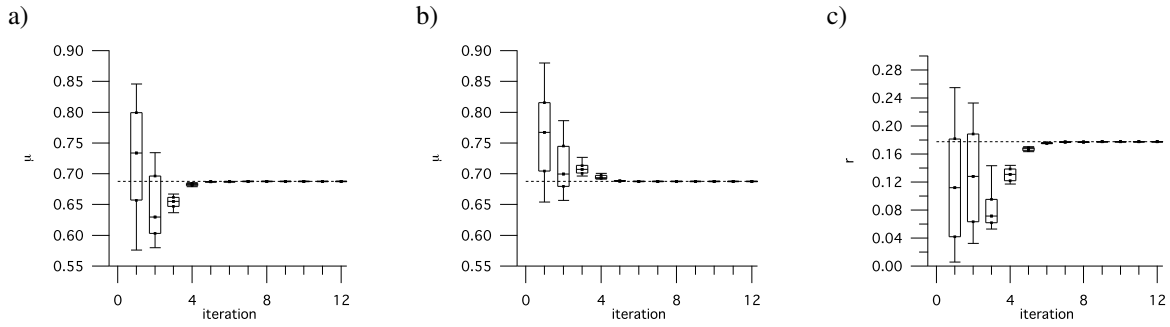

Figure 5: a) mutuality obtained with Eq. 7, b) mutuality obtained with Eq. 8 and c) degree-correlation of the sampled networks.

3.3 Closeness and Betweenness

This section presents two further network statistics, closeness and betweenness. These statistics are defined in a non-local manner, which complicates the previously applied correction of the snowball-sampling induced degree bias. While an important aspect of our ongoing research is the design of a proper bias correction logic even for these statistics, in this article, we constrain ourselves to an illustration of these statistics' basic features and an application of the hitherto deployed local correction logic.

The closeness of a vertex is its mean geodesic distance to all other vertices in the network, i.e., the average shortest path length. Closeness is ill-defined on disconnected networks, which is the case for snowball-sampled networks where the seed vertices' ego-centric networks did not yet connect to a giant component. However, we can calculate the closeness only for the vertices within a connected component. We denote the closeness of a vertex v as CC_v (closeness centrality) which is defined as

$$CC_v = \frac{\sum_{w \in \mathcal{V}' \setminus \{v\}} d_G(v,w)}{n' - 1} \tag{10}$$

where $\mathcal{V}'$ is the set of all vertices reachable from v, $n' \geq 2$ the number of elements in $\mathcal{V}'$, and $d_G(v,w)$ the shortest path from vertex v to w. The mean closeness of the network is calculated by averaging CC_v over all sampled vertices**.**

The closeness of the sampled networks increases with each iteration, i.e., with increasing system size (Fig. 6a). We can assume that the closeness values will increase if the seed vertices' ego-centric networks start to connect to larger components. If the ego-centric components connect the amount of distant vertices escalates. This would explain, why the closeness already increases to approximately four in the first iteration. Note the there are some simulation runs in which the average closeness is greater than one even in the 0th iteration. In these runs some sub-networks already connect in the 0th iteration (via anonymous vertices).

Figure 6c depicts the correlation between closeness and degree. Vertices with higher degrees tend to be closer compared to vertices with lower degrees, which is intuitively plausible. In this regard, we apply the weighting approach as done for degree and clustering also to closeness

$$\langle CC \rangle^{(i)} = \frac{\sum_{v \in \mathcal{V}(0\ldots i)} CC_v / P^{(i)}_{k_v}}{\sum_{v \in \mathcal{V}(0\ldots i)} 1 / P^{(i)}_{k_v}} \tag{11}$$

However, the results show only improvements of minor magnitude (Fig. 6b).

Betweenness is an other centrality measure of a vertex within a network and tells us something about the importance of a vertex. The betweenness centrality of a vertex v is defined as

$$BC_v = \sum_{s \neq v \neq w, s \neq w} \frac{\sigma_v(s,w)}{\sigma(s,w)} \tag{12}$$

where $\sigma_v(s,w)$ denotes the number of shortest path between vertex s and w that pass through v and $\sigma(s,w)$ denotes the total number of shortest paths between s and w. The mean betweenness of the network is estimated by averaging BC_v over all sampled vertices.

Figure 6d depicts the mean betweenness values for the arxiv.org network. The picture resembles Fig. 3a of the mean degree where the values are overestimated in the first iterations. Figure 6f shows that there is also a

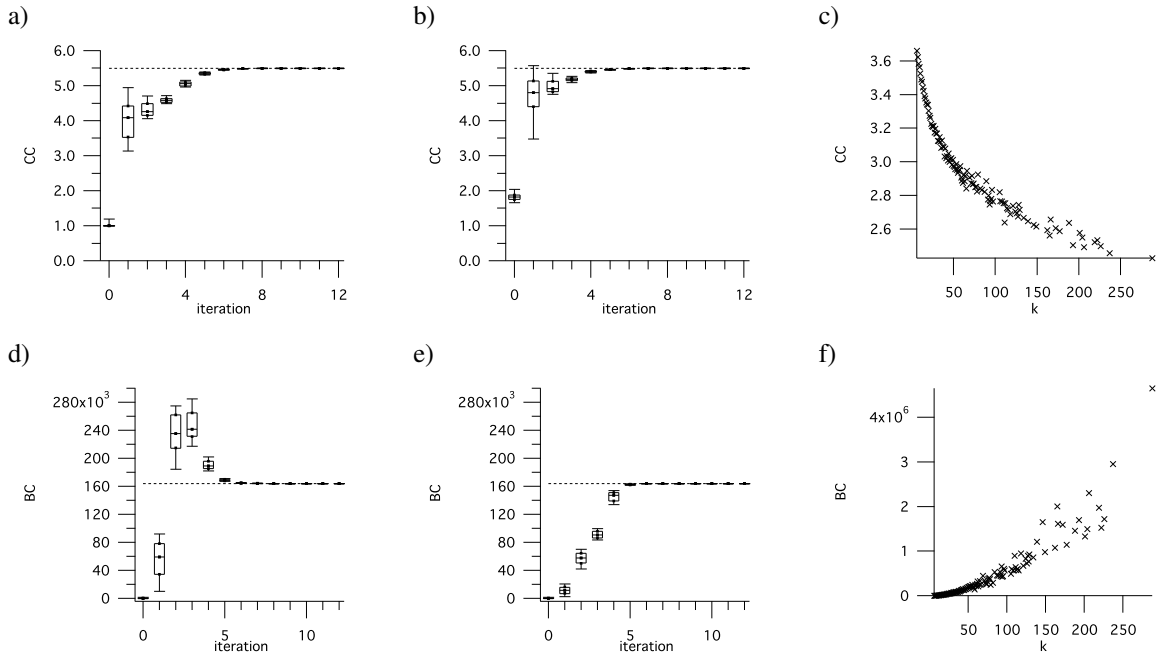

Figure 6: Centrality measures for the arxiv.org network. a) closeness b) closeness with probability weighting (Eq. 11) c) closeness-degree-correlation, d) betweenness, e) betweenness with probability weighting and f) betweenness-degree-correlation.

betweenness-degree-correlation. Vertices with higher degree tend to have a higher betweenness while vertices with low degree tend to have lower betweenness values. For Barabási-Albert-type[1] networks, this correlation has been shown analytically [4], while for assortative networks (such as the arxiv.org network), the analytical solution only holds for vertices with low degrees [5]. Figure 6e depicts the results when probability weighting is applied which leads to the expression

$$\langle BC \rangle^{(i)} = \frac{\sum_{v \in \mathscr{V}(0\ldots i)} BC_v / P_{k_v}^{(i)}}{\sum_{v \in \mathscr{V}(0\ldots i)} 1 / P_{k_v}^{(i)}}. \tag{13}$$

Figure 6e shows now a completely different course. The betweenness appears to scale nearly linearly with the number of iterations.

4 Discussion

In this article, we pointed out an issue that arises when conducting a snowball sampling on social networks. The snowball sampling technique over-represents vertices with high degrees in early iterations. As a consequence, the estimate of the degree distribution and all statistical properties that correlate with the degree distribution are biased. We introduced a method to compensate this bias. Our proposed estimator of the sampling probability can be easily calculated since it only depends on the considered degree and on the number of sampled vertices in the previous iteration. We showed that this approach to compensate the bias leads to reasonable results in the estimation of the mean degree and the clustering coefficient. However, further investigations of bias-corrections for measures of closeness and betweenness are necessary that account more properly for the nature of these statistics.

Furthermore, an open question is if the results obtained with the specific social network used in this work can be applied to other social networks. The available large-scale network data often represents some kind of collaboration network, e.g., among authors or actors, which only is one specific type of social network. Two further issues should be addressed in further research: (i) how does the number of initial seed vertices affects the sampling process, and (ii) what happens if some vertices are non-responding, i.e., respondents reject to participate in the survey.

[1] Barabási-Albert networks are models of networks with preferential attachment and follow a power-law degree distribution.

5 Acknowledgement

This work was funded by the VolkswagenStiftung within the project “Travel impacts of social networks and networking tools”.

References

[1] UCINET, Accessed July 2008. URL `http://www.analytictech.com/ucinet/ucinet.htm`.

[2] Vladimir Batagelj and Andrej Mrvar. Pajek: Program for Analysis and Visualization of Large Networks, Accessed July 2008. URL `http://pajek.imfm.si/`.

[3] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *E-Print*, arXiv:0706.1062v1, 2007.

[4] K.-I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks. *Physical Review Letters*, 87(27), 2001.

[5] K.-I. Goh, E. Oh, B. Kahng, and D. Kim. Betweenness centrality correlation in social networks. *Physical Review E*, 67(017101), 2003.

[6] Leo A. Goodman. Snowball sampling. *The Annals of Mathematical Statisitics*, 32(1):148–170, 1961.

[7] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1), 2008.

[8] G. Kossinets. Effects of missing data in social networks. *E-Print*, arXiv:cond-mat/0306335v2, 2008.

[9] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(016102), 2006.

[10] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 98:404–409, 2001.

[11] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20), 2002.