Using snwoball sampling to measure the “degrees of separation”

Johannes Illenberger[a]

[a]*Berlin Institute of Technology, Group of Transport Systems Planning and Transport Telematics, Salzufer 17–19, D-10587 Berlin, Germany, +49 30 31478793, illenberger@vsp.tu-berlin.de*

Abstract

This article addresses the application of snowball sampling as a mechanism to estimate the average shortest path length of a social network. In several simulation experiments the estimates of the average shortest path length depending on the number of seeds with which the snowball is initialised and the response-rate are investigated. The results are compared to a simple node sampling design. From the simulation results it is concluded that there is no rule of thumb how to obtain precise estimates since the progress of the snowball mechanism is governed by the underlying network topology. Yet, a researcher might consider the following aspects: (i) the quality of estimates depends only on the sample size and is independent from the number of snowball iterations conducted and the response rate, (ii) snowball sampling performs better than node sampling, and (iii) initialising the snowball with a lot of seeds induces a bias towards longer paths.

Keywords: snowball sampling, average shortest path length, simulation, social networks

1. Introduction

In 1967 the american psychologist Stanley Milgram conducted an experiment in which he asked a randomly selected set of persons to send a message to a preselected target person [1]. If a person did not know the target person directly, she should pass the message to an acquaintance she knew on a first-name basis and which she believed is more likely to know the target person. The experiment yielded an average of 5.4 intermediates a message took to reach the target person. From this result, Milgram made inference about the number of edges connecting two randomly selected persons in a social network. His hypothesis became famous as the “six degrees of separation”.

Considering that the above hypothesis is based on a sample size of only 44 completed chains, the question arises whether there is an evidence that everyone is connected to each other by only six edges. Extensive research on social networks constructed from proxy-data, such as email archives [2], databases of scientific articles [3], or movie actor databases [4], revealed that most of them exhibit a surprisingly low number of edges connecting a random pair of vertices and thus fortify Milgram’s hypothesis. Yet, these networks are usually embedded in some institutional context and thus differ from Milgram’s first-name-based acquaintance network. To collect data on a real-world acquaintance network one usually needs to turn to rather traditional sampling strategies, that is, directly asking respondents to disclose their social contacts. Since this process involves a lot of effort, the obtained social networks are often incomplete and any inference of network properties requires some (more or less) elaborated estimation technique.

Sampling designs for social networks discussed in sociology can be roughly categorised into ego-centric sampling designs and link-tracing designs. In the ego-centric sampling approach, a random set of respondents, so-called *egos*, is asked to report their social contacts, denoted as *alters*. This design is straightforward to implement and, as long as the selection of egos is independently and identically distributed, requires no elaborate estimation technique. A link-tracing design, commonly also denoted as *snowball sampling*, collects new respondents by tracing links form already sampled egos. In contrast to ego-centric sampling, snowball sampling allows to reveal network structures that go beyond first degree relations. However, since snowball sampling is an adaptive sampling design, that is, the selection of new respondents depends on already existing respondents, the inference of network characteristics requires an appropriate estimation methodology.

Snowball sampling has found its application in a variety of domains. Probably the most common known application represents *Respondent-Driven Sampling* [5, 6], which is often used to access hidden populations. Snowball sampling, however, can also be used to reveal the topology of a social network [7, 8]. There is plenty of research that addresses the estimation network characteristics, such as size, degree distribution, transitivity, or assortativity [9–14]. Albeit global network properties, specifically the average shortest path length, represent decisive network attributes to characterise the spreading of information and infectious diseases [15], their estimation has gained just little interest [16–18].

In this paper we investigate in snowball sampling as a mechanism to measure the average shortest path length of a social network. We conduct several simulation experiments on public available social network data sets and investigate in the measurements of the average shortest path length with respect to the sample size and the sampling design. Additionally, we compare the results with a simple ego-centric sampling design. The study is targeted at real-world application and considers incomplete networks caused by non-responding persons. The remainder of this article is organised as follows: In Sec. 2 we introduce the considered snowball sampling design and the ego-centric sampling design, which is used as a reference design. Studies that address similar problems are presented and discussed. Simulation experiments with both sampling designs are presented in Sec. 3. Section 4 closes the paper with a conclusion that highlights some aspects a research might consider when conduction a snowball sample.

2. Sampling designs

2.1. Snowball sampling

Snowball sampling is an iterative sampling design that can be implemented in quite different variants. Variants commonly known in computer science are *random walks*, *breadth first search*, or *forest fire* algorithms. A well known implementation in sociology represents Respondent-Driven Sampling [5, 6], which is often used to reveal hidden or hard-to-reach populations (for instance drug-injection users or HIV-infected people [19, 20]). In this study we use snowball sampling as mechanism to reveal the topology of the social network. It is assumed that the sampling frame is given, this mean that the population from which the samples are drawn is known to the researcher. The sampling mechanism proceeds as follows:

Draw an identically and independently distributed set of initial respondents, so-called *seeds*, and ask them to report their social contacts which are then invited to participate in the survey as well. The new respondents are asked to report their social contacts which are in turn also invited. This iterative process is continued until no new social contacts are named or until the desired number of samples is collected. A person my reject to participate the survey. It is then denoted as non-responding. Sampling is done without replacement, that is, a person is never inquired multiple times, even if she is non-responding. Referring to the nomenclature in sociology, we denote respondents (persons who filled out a questionnaire) as *egos* and persons who have been named by egos but who did not participate in the survey as *alters*. Consequently, alters are those person who are either non-responding or who where indicated by egos but have not been inquired because the sampling process terminated before. This sampling design is equivalent to breadth-first search.

One can consider Milgram's experiment as a variant of this snowball sampling design. It differs from this mechanism in its *branching rule*. This mean that instead of reporting all social contacts, in Milgram's experiment a respondent is asked to report only one contact: Namely the one she thinks is most likely to be able to forward the message to the target person.

2.2. Ego-centric sampling

From an operational perspective, ego-centric sampling (also denoted as node sampling) is much simpler compared to snowball sampling: Draw a random set of independently and identically distributed respondents (egos) and ask them to report their social contacts (alters). One then obtains a set of so-called ego-centric networks with the alters positioned in a star-like alignment around the ego. Like in snowball sampling, respondents may reject to participate in the survey and the sampling process is without replacement. Yet, alters can be named by different egos. Statistical inference from the sample is straightforward

Table 1: Properties of test networks. N = number of vertices, $\langle k \rangle$ = mean degree, C = local clustering coefficient [25], ℓ = average path length.

Description	N	$\langle k \rangle$	C	ℓ	Alias
Collaboration network, astro physics [3]	16'706	14.5	0.64	4.8	*astro-ph*
Collaboration network, condensed matters [3]	40'421	8.7	0.64	5.5	*condmat*
Who-trusts-whom network of Epinions.com [26]	75'879	10.7	0.14	4.16	*epinions*
Email network from a EU research institution [27]	265'214	2.7	0.07	4.1	*email*
Collaboration network, high energy physics theory [3]	8'361	3.8	0.44	7.0	*hep-th*
Slashdot social network [28]	82'168	12.3	0.06	4.07	*slashdot*

because it is unbiased. Unless ego-centric networks to not overlap, which is the case if multiple egos name the same alter or two egos name each other, the analysis of network properties is restricted to first-degree relations.

2.3. Related Work

Snowball sampling is extensively discussed in sociology. The discussion covers the application of snowball sampling to reveal hidden populations [10, 20, 21] as well as the application to uncover the topology of the network [7, 13, 16, 22]. Considering the estimation topological properties, existing literature focuses on the estimation of rather local properties, such as size, degree distribution, degree correlation, and transitivity. There is only little research addressing the application of snowball sampling as well as ego-centric sampling to make statements about the average shortest path length [16–18].

Lee et al. [16] investigate the estimates of the average path length on snowball sampled networks. In their implementation of snowball sampling only one seed-vertex is randomly drawn and all neighbours are recruited during the expansion step. Amongst others, they conduct numerical simulations on a collaboration network and find that the average path length increases with the sample size. A reasonable estimate of the true value, however, is only archived if nearly the entire network is sampled. The results of Lee et al. are reproduced by Ye et al. [18] using the same sampling design but different test networks. Ebbes et al. [23] compare vertex sampling, edge sampling and the snowball sampling variants random walk, forest fire and breath first search. They show that the snowball sampling designs generally perform better than vertex sampling and edge sampling, but differ with respect to the degree distribution of the test network. However, they make no precise statements about the behaviour of the average path length estimates regarding the sample size. A couple of studies discuss the application of random walks to reveal global network properties [17, 24]. The nature of random walks, however, lend themselves to the automatic processing of online networks, rather than to the manual sampling of real-world acquaintance networks.

The above discussed studies provide useful insights into the dynamics of snowball sampling. Yet, their applicability to real-world studies is limited. First, in a real-world study, the researcher is faced with non-response effects, which especially when using snowball sampling hinders the expansion process. Second, the network does not necessarily consist of one connected component. When using snowball sampling, selecting just one seed bears the danger of being trapped in an isolated subgraph. Therefore, real-world applications usually initialise the snowball with multiple seeds. The snowball sampling design used in this study addresses both aspects.

3. Simulation Experiments

3.1. Setup

The experiments are conducted with a Monte Carlo simulation implementing the snowball sampling design as well as the ego-centric sampling design as a reference design. In case of the snowball sampling design, the simulation creates a snapshot of the hitherto sampled network each time a *bridge* is detected.

A bridge is defined as a link that connects two nodes tracing back their recruitment-chain would lead to distinct seeds. Although this event is not the only change to the sampled network that causes a change of path length properties, the resulting snapshot intervals feature a sufficient resolution of the evolving of the sampled network. The operation requires no additional computation time because it is kept track of the recruitment-chain already within the expansion process. In case of the ego-centric sampling design, snapshots are created after sampling n additional egos until N egos have been sampled in total, where the following configurations are used: $n = 10$ for $N < 200$, $n = 100$ for $N < 1'000$, $n = 1'000$ for $N < 10'000$, and $n = 5'000$ for $N \geq 10'000$. Creating and analysing a snapshot each time a node is sampled would be computationally infeasible.

Each snapshot is analysed regarding the average shortest path length. The true average shortest path length, ℓ, is defined as the average number of edges of all shortest paths between all vertices. The average path length, $\hat{\ell}$, calculated on a *snapshot* from the snowball simulation considers only the shortest paths between seeds. Paths with $l = \infty$, which means that there exists no path between both seeds, are removed from the calculation. There is no definition of seeds in the ego-centric design. To obtain a compatible measurement anyway, the calculation on an ego-centric sampled network considers the shortest paths between the first s sampled egos, where s is the number of seeds in the snowball sample. The remainder of this article often discusses the distribution of shortest paths, containing "shorter" and "longer" shortest paths. To avoid confusion, we introduce the convention that a "path" refers to the shortest path in the current snapshot and the shortest path in the original network is denoted as the "true shortest path".

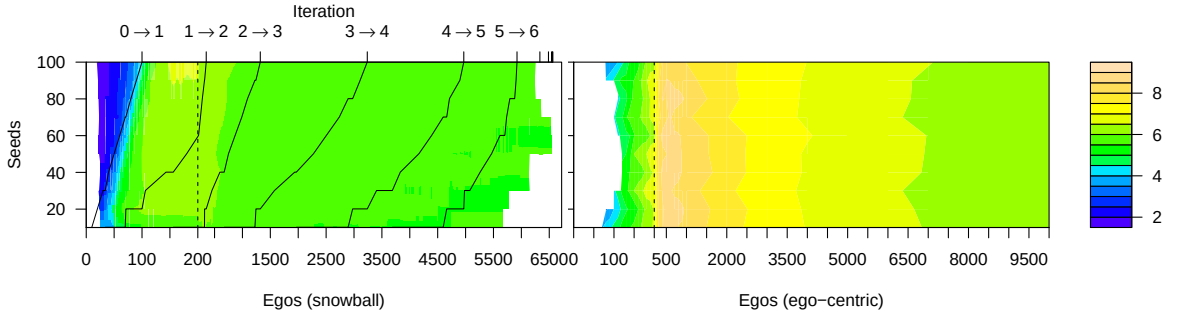
At the beginning of each simulation experiment a fraction of $1-\alpha$ of nodes are marked as non-responding. The effects of the response rate α is investigated by varying this parameter from 0.1 to 0.5 in 0.1 steps. Due to long computation times configurations with response rates above 0.5 are not treated. This is reasonable considering that response rates above 0.5 are unlike to be observed in real-world applications. The number of seeds is varied from ten to 100 in steps of ten. In the ego-centric sampling design, this means that the number of nodes used for the calculation of the average shortest path length is varied. Each parameter configuration is repeated 200 times with different random seeds. Averages over the simulation ensemble are calculated if at least 50 simulations yielded a valid measurement.

The simulation studies consider six social networks. Three networks represent co-authorships of scientific papers generated from the arxiv.org database, two networks are generated from user interactions of the websites slashdot.org and epinions.com, and one network is generated from email communications of a European Union research institution. Table 1 lists the network properties.

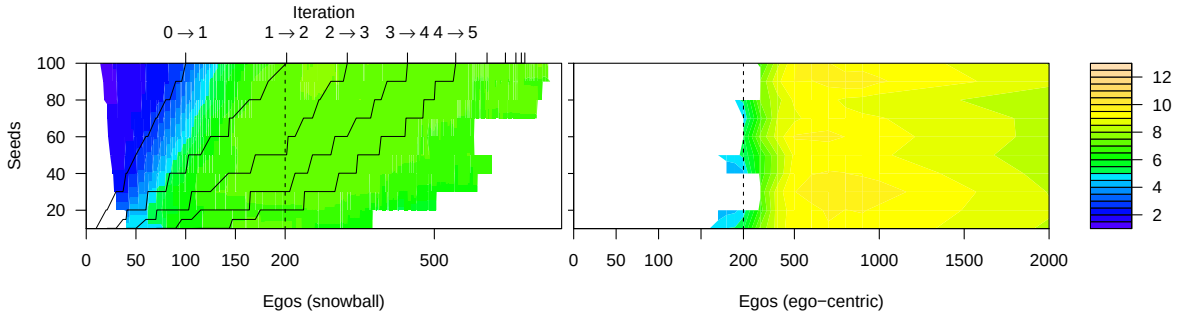
3.2. Measuring the average path length

Figure 1 (left) exemplarily shows the average shortest path length $\hat{\ell}$ estimated on the snowball samples of the *condmat* network, *hep-th* network, and *slashdot* network depending on the number of egos and the number of seeds. The measurements of $\hat{\ell}$ in the first snapshots yield values of $\hat{\ell} < 2$. The values then increase with the growth of the sample size, however, overestimate the true average shortest path length before they level off at ℓ. Clearly observable is a correlation between the number of seeds and the measurements of $\hat{\ell}$: First, if the snowball is initialised with more seeds, it requires more egos to approximate ℓ. Second, if the number of seeds is increased, the overestimation of ℓ becomes more pronounced. Generally, this observation holds for all considered networks. The extend of the bias, however, differs between the networks: The bias in the *hep-th* (Fig. 1(b)) network is less pronounced compared to the *condmat* network (Fig. 1(a)).

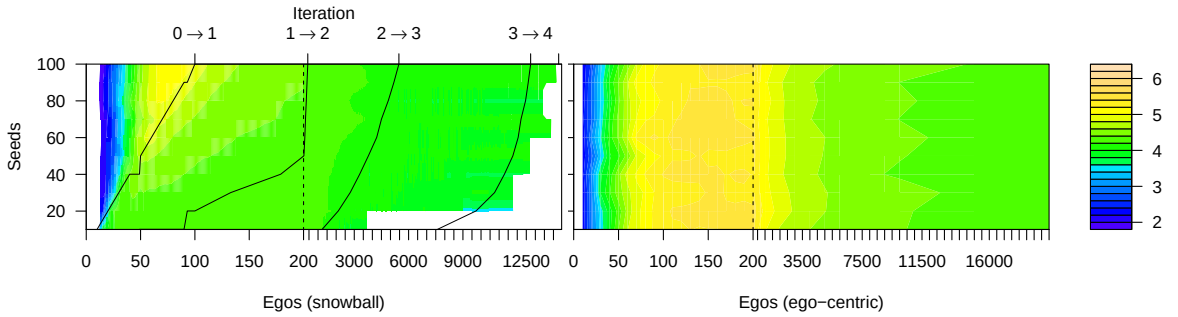
The dynamics of the snowball can be explained as follows: Values of $\hat{\ell} < 2$ in the first snapshots indicate that in some simulations bridges are already detected during the drawing of seeds. That is, seeds are adjacent to each other so that $l = 1$. The probability of such configurations increase if the fraction of seeds increases. With the progress of the snowball, the components emerging out of the seeds start to connect to greater components. Additionally to those paths that initially exist if two seeds are adjacent, first paths between seeds with $l > 1$ are detected. Among the distribution of paths in the original network, the hitherto detected paths are those with shorter length. This means that the snowball does not reveal the paths in random order but in ascending order regarding their length. This explains why it requires more egos with a greater number of seeds, compared to a smaller number of seeds, to obtain the same value of $\hat{\ell}$: If the snowball is initialised with more seeds, more possible paths exist, and consequently the absolute number of



(a) *condmat*



(b) *hep-th*



(c) *slashdot*

Figure 1: Average path length $\hat{\ell}$ (indicated by colour) over number of egos for a configuration with $\alpha = 0.3$. Colours are adjusted so that green colour indicates the true average shortest path length ℓ. Note the change of the x-axis scaling at 200 egos.

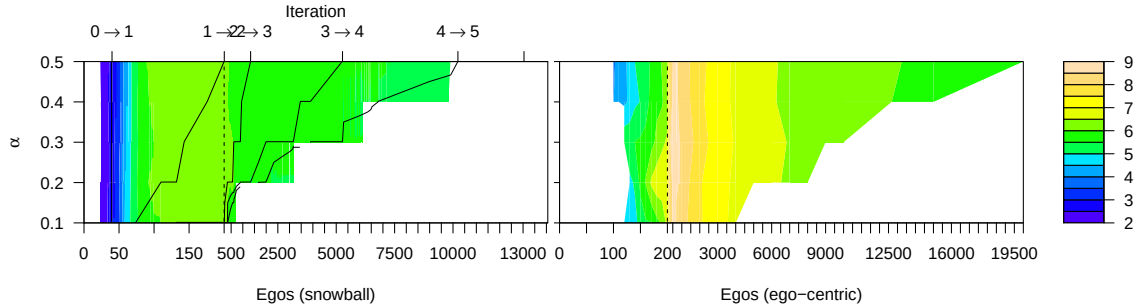


Figure 2: Average path length $\hat{\ell}$ (indicated by colour) over α and the number of egos for the *condmat* network with 40 seeds. Colours are adjusted so that green colour indicates the true average shortest path length ℓ. Note the change of the x-axis scaling at 200 egos.

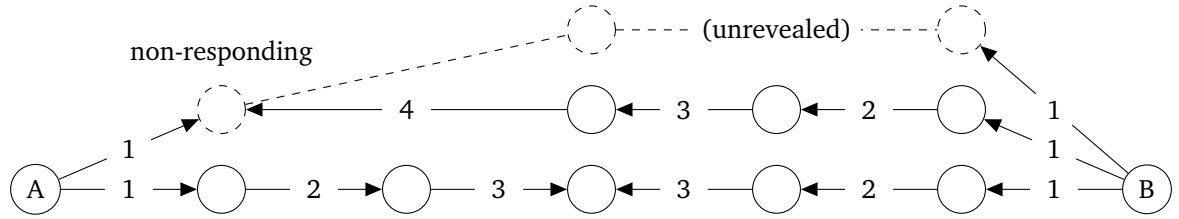


Figure 3: A seed-pair connected by three paths with $l = 6$, $l = 5$, and $l = 4$. Because there is a non-responding node in the path with $l = 5$, this path takes one iteration more (4 iterations) to be revealed than the longer path with $l = 6$ (3 iterations). The upper path ($l = 4$) remains unrevealed.

shortest paths is greater. Since the sampling algorithm detects the paths in ascending order it terms of path lengths, it first detects all the short paths before it detects the long paths. That is, the sampling algorithm needs to detect more paths, and thus more nodes, to have a sufficient number of long paths so that the same value of $\hat{\ell}$ is obtained.

The overestimation of ℓ is caused by so-called "indirect paths". Consider three seeds A, B, and C. In the first iteration, the components that emerge out of the seeds connect through a common neighbour. Hence, there is a path from A to B with $l_{A\to B} = 2$ and a path from B to C with $l_{B\to C} = 2$. Because A is connected to B and B is connected to C there also exists an indirect path from A to C over B with length $l_{A\to C} = 4$. At this point it is unknown if this path is the true shortest path or if there exists a shorter path with $l_{A\to C} = 3$. If the latter is true, that path will be revealed in the succeeding iteration. Consequently, the appearance of indirect paths induces a systematic bias towards longer paths. Moreover, this also means that situations may occur where the number of detected paths remains constant but their length decreases with the progress of the snowball. An important implication of this observation is that the number or share of detected paths is not a valid measure to make statements about the reliability of the estimates. The intensity of the systematic bias depends on the number of seeds. If more seeds are drawn, the probability of indirect paths to occur increases. Additionally, the probability that more than three components connect increases, too, and consequently longer indirect paths that go over multiple seeds are possible.

Omitting indirect paths during the statistical analysis does not improve the estimates of $\hat{\ell}$. There is no way to distinguish between an indirect path that will be later replaced by a direct path or an indirect path that is in fact the true shortest path. The consequence is that this approach would reduce the bias for small sample sizes, where the share of indirect paths is likely to be dominating, but would increase the bias for larger sample sizes where the remaining indirect paths that are in fact true shortest paths are omitted.

Figure 1 (right) shows the average path length $\hat{\ell}$ estimated on the ego-centric sampled networks. Similar to the snowball sampled networks, the average shortest path length is first underestimated with small sample fractions, then overestimated, and finally levels off at the true value with large sample fractions. The explanation is analogous to above: With small sample fractions there are some ego-centric networks that are next to each other, that is, egos have a common neighbour or name each other. With increasing sample fractions the density of ego-centric networks increases so that paths, most likely indirect paths, that go over multiple egos are detected. Just with large sample fractions the density of ego-centric networks is that high so that the direct paths are revealed.

In contrast to the snowball sampling design, the number of egos required until the first path are detected is at average greater and correlates with the true average shortest path length. The *slashdot* network exhibits the lowest average shortest path length ($\ell = 4.07$) and the first paths between seeds are discovered with just ten egos (Fig. 1(c)). About 200 egos are required to detect paths in the *hep-th* network, which exhibits the greatest average path length ($\ell = 7.0$, Fig. 1(b)). The *condmat* network is in between (Fig. 1(a)). Moreover, the ego-centric sampled networks show no correlation between $\hat{\ell}$ and the number of seeds. This indicates that it makes no difference if $\hat{\ell}$ is calculated on the basis of ten end-points or 100 end-points.

Considering the snowball sampling design, a second source of bias is related to the response rate. Consider two seeds A and B, which are connected through two paths: A path of length $l = 6$ and the true shortest path with length $l = 5$. If the recruitment chain that follows the true shortest path dies at seed A's

first expansion step because the alter does not responds, then this path can only be closed from the direction of seed B. This takes four iterations. If along the second path with $l = 6$ all nodes are responding, then the snowball requires three iterations to close this path. Hence, the true shortest path is detected one iteration after the path with $l = 6$ is detected. This effect also biases the estimates of $\hat{\ell}$ towards longer paths (Fig. 3).

The estimates of $\hat{\ell}$ show almost no sensitivity towards variations of the response rate α. Figure 2 shows $\hat{\ell}$ over α and the number of egos exemplarily for the *condmat* network. The other networks show the same behaviour. This is rather unsurprising considering the ego-centric sampling design. Yet, even with the snowball sampling design, it is only the number of egos and not the number of conducted iterations that determines the estimates of ℓ.

3.3. Confidence of measurements

The findings of the previous section provide a useful insight into the dynamics of a snowball sample. Yet, considering a real-world application the observations are of limited practical use because they represent an average over a sample ensemble. In a real-world application repeated sampling is usually prohibitively expensive. It is thus useful to determine a reliability measure of a single estimate. Therefore, the absolute value of the relative error

$$\epsilon(\hat{\ell}) = \left| \frac{\hat{\ell} - \ell}{\ell} \right| \tag{1}$$

is calculated for each simulation run, and the 0.8 percentile of $\epsilon(\hat{\ell})$ is determined. This means that if the 0.8 percentile of $\epsilon(\hat{\ell})$ is 0.1, 160 of 200 simulation runs do not exceed an error of $\epsilon(\hat{\ell}) = 0.1$. Or one could say: with a certainty of 80 % the estimate $\epsilon(\hat{\ell})$ does not exceed an error of 0.1.

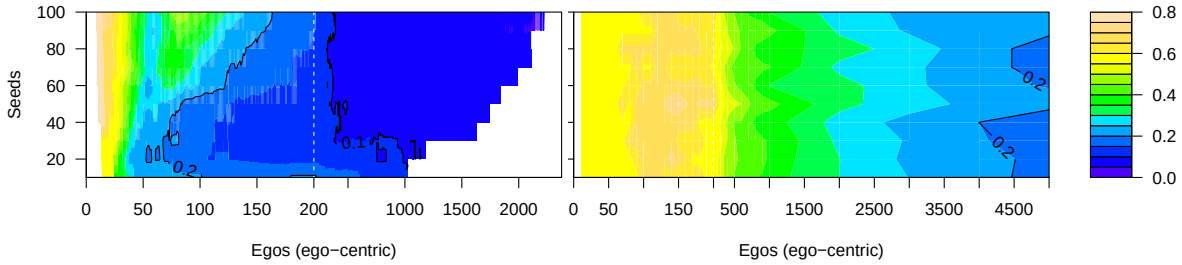
Figure 4 shows the 0.8 percentile of $\epsilon(\hat{\ell})$ over the number of seeds and egos. The response rate α is set to 0.3. Generally, the snowball sampling design performs significantly better than the ego-centric sampling design, that is, it requires less egos to obtain the same error. In none of the plots the ego-centric design is able to fall below an error of 0.1 even though the maximum possible sampling fraction is higher. Using the ego-centric design, the maximum number of egos that can be collected is $n \approx \alpha N$. Non-responding nodes may hinder the snowball to percolate into new regions of the network. Thus, the maximum number of egos that can be collected with the same response rate is less in the latter design.

Considering the snowball sampled *hep-th* network (Fig. 4(d)) one requires at least about 60 seeds in order to achieve an error of less than 10 % at all. If too few seeds are drawn the snowball dies (due to the low response rate) before it is able to sample a sufficient number of egos. Similar effects are observed with the *astro-ph* network (Fig. 4(a)) and the *epinions* network (Fig. 4(c)), however, with a much lower threshold of about 20 seeds. The snowball sampled *email* network (Fig. 4(b)) represents the only network with which it is possible to obtain a relative error of less than 10 % with just 10 seeds. The lowest number of egos required to fall below a 10 % error is achieved with the *epinions* network. For a configuration with 30 seeds it requires only to sample 50 egos in order to fall below a 10 % error. Generally, the results reflect the observations of the previous section: When using fewer seeds in a snowball sample, fewer egos need to be sampled in order to obtain the same estimation quality. However, there appears to be a lower threshold at about 30 seeds where the share of dying recruitment chains is too high to obtain reasonable estimates.

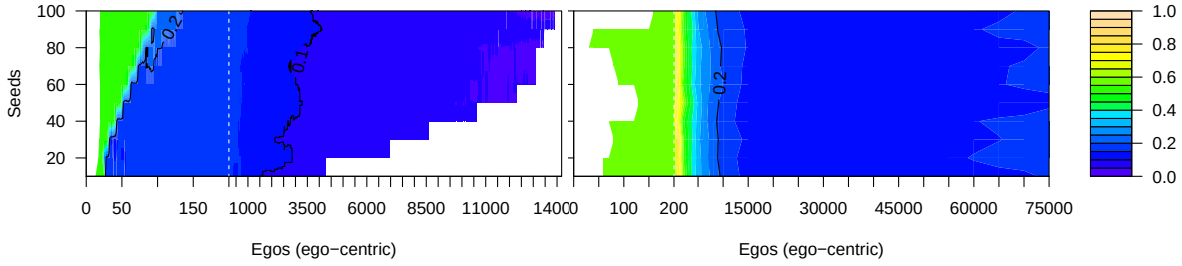
For snowball configurations with at least 30 seeds, moderate errors, say less than 20 %, are achievable with quite few vertices, less than 200 egos for all three networks. One may say that reasonable results can be obtained with rather few seeds and few egos. To obtain more precise results, say an 0.8 percentile of $\epsilon(\hat{\ell}) < 0.1$, one requires to collect significant more egos. The results of the ego-centric sampled networks vary too much so that such a rule of thumb cannot be proposed.

3.4. Impacts of the network topology

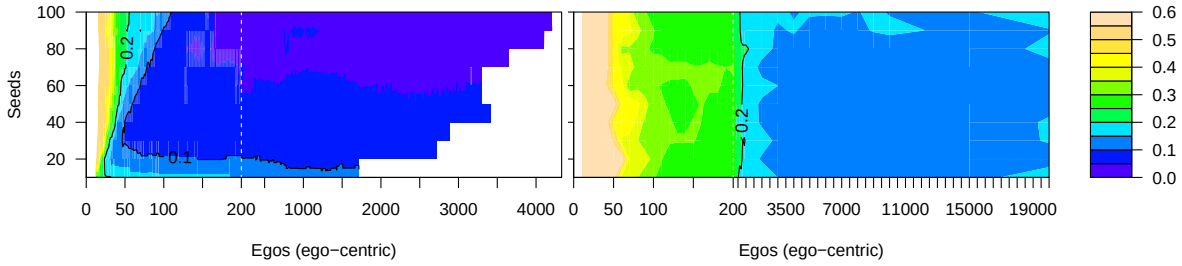
An interesting question is how the topology of the network influences the estimation error of $\hat{\ell}$. This is investigated using a simulation configured with 40 seed-vertices and $\alpha = 0.3$. For each test network the number of egos required to fall below the 0.8 percentile of the 10 %, 20 %, and 30 % relative error $\epsilon(\hat{\ell})$ is



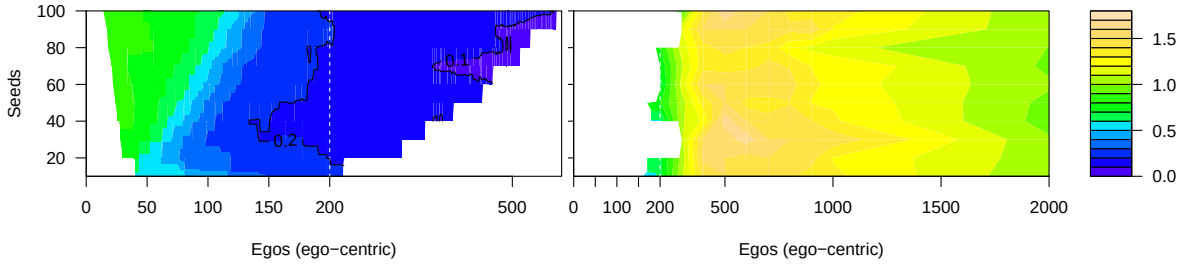
(a) *astro-ph*



(b) *email*



(c) *epinions*



(d) *hep-th*

Figure 4: 0.8 percentile of the relative error $\epsilon(\hat{\ell})$ over number of egos for a configuration with $\alpha = 0.3$. Note the change of the x-axis scaling at 200 egos.

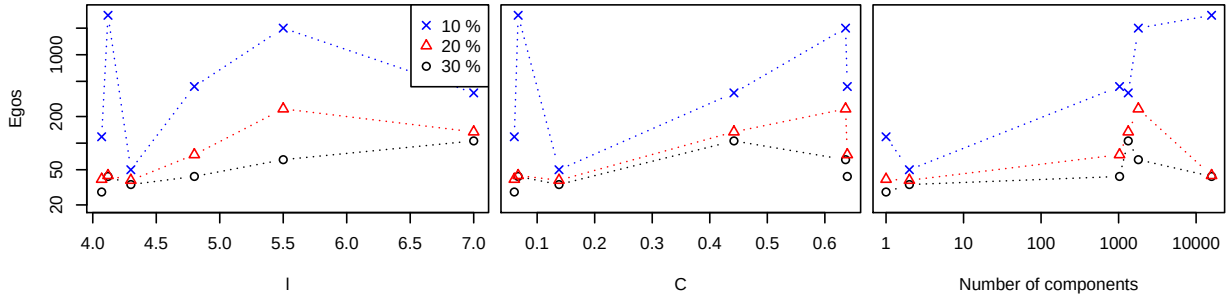


Figure 5: Number of egos required to fall below the 10 %, 20 %, and 30 % relative error $\epsilon(\hat{\ell})$. Values plotted over network properties average path lentgh, local clustering coefficient, and number of isolated components. Snowball configuration with 30 seeds and $\alpha = 0.3$.

determined. The obtained number of egos is then plotted against the networks' true average path length, local clustering coefficient, and number of isolated components.

The results for snowball sampled networks (Fig. 5) are summarised as follows: If the average path length, local clustering coefficient, or the number of isolated components increases, the number of required egos increases, with some exceptions, too. Regarding the average path length, it is quite intuitive that it requires more egos to detect longer paths. A bit off is the *email* network which is assumably because of its high fragmentation (a lot of smaller isolated components). When the clustering coefficient increases, vertices become locally well connected but at the costs of long paths to other network regions (a bit off are again the *email* network and the *astro* network). With the increasing fragmentation of a network, it becomes more probable that seeds are placed in smaller isolated components. These seeds are never able to connect to other seeds placed in other components. Thus, nodes are sampled but without ever contributing to the detection of paths between seeds.

Considering the ego-centric sampled networks, in none of them the error falls below 0.1 and just for some networks it falls below 0.2 and 0.3. No clear correlation between network properties and required sample size is observed. This indicates that the trends observed with the snowball design are a result of the very sampling design rather than a result of the network structure.

4. Conclusion

This paper discusses the application of snowball sampling to estimate the average shortest path length in a social network. While it does not propose a rule of thumb how to obtain a precise estimate without sampling thousands of nodes, it highlights some aspects a researcher might consider when conducting a snowball sample.

First, there is no precise functional relation between the estimation error and the sample size. How many egos need to be sampled depends on the network topology, which in turn is the property of interest. Regarding this dependency, the following rule approximately holds: If the average path length, the clustering coefficient, or the number of isolated components increases the required sample size increases, too.

Second, the response rate has no impact on the estimates of the average shortest path length. This is an important observation considering that the response rate is usually out of the researcher's control. Further, this indicates that the number of conducted snowball iterations does not affect the estimation quality given a fixed sample size.

Third, one should initialise the snowball with just a few seeds, where 30 seeds appears to be some minimum value. This means that fewer seeds and more snowball iterations are to be preferred over more seeds and fewer snowball iterations in order to sample the same number of nodes. Using more seeds bears the risk of a bias towards longer paths induced by indirect paths. One may even artificially hold the response rate low (by not expanding all egos) in order to do more snowball iterations.

The comparison with the ego-centric sampling approach shows that snowball sampling requires significantly smaller sample fractions in order to make reasonable statements about the average shortest path length. Even though snowball sampling involves more effort from an operational perspective and is biased when estimating other properties (for instance the degree distribution), it is the preferred sampling design to reveal the global structure of a social network. Finally, measuring the average shortest path length in a social network remains still challenging or probably even impossible. Yet, the snowball sampling approach provides some evidence about the order of magnitude of the "degree-of-separation" distribution.

5. Acknowledgement

We thank Kai Nagel for helpful suggestions and support. This work was funded by the VolkswagenStiftung within the project "Travel impacts of social networks and networking tools".

References

[1] S. Milgram, The small world problem, Psycology Today 2 (1967) 60–67.
[2] H. Ebel, L.-I. Mielsch, S. Bornholdt, Scale-free topology of e-mail networks, Physical Review E 66 (3) (2002) 1–4.
[3] M. E. J. Newman, The structure of scientific collaboration networks, Proceedings of the National Academy of Sciences of the United States of America 98 (2) (2001) 404–409.
[4] L. A. N. Amaral, A. Scala, M. Barthélémy, H. E. Stanley, Classes of small-world networks, Proceedings of the National Academy of Sciences of the United States of America 97 (21) (2000) 11149–11152.
[5] D. D. Heckathorn, Respondent-driven sampling: A new approach to the study of hidden populations, Social Problems 44 (2) (1997) 174–199.
[6] D. D. Heckathorn, Respondent-driven sampling II: Deriving vaild population estimates from chain-referral samples of hidden populations, Social Problems 49 (1) (2002) 11–34.
[7] C. Wejnert, Social network analysis with respondent-driven sampling data: A study of racial integration on campus, Social Networks 32 (2010) 112–124.
[8] J. Illenberger, M. Kowald, K. W. Axhausen, K. Nagel, Insights into a spatially embedded social network from a large-scale snowball-sample, European Physical Journal B 84 (4) (2011) 549–561. doi:10.1140/epjb/e2011-10872-0.
[9] L. A. Goodman, Snowball sampling, The Annals of Mathematical Statistics 32 (1) (1961) 148–170.
[10] O. Frank, T. Snijders, Estimating the size of hidden population using snowball sampling, Journal of Official Statistics 10 (1) (1994) 53–67.
[11] E. Volz, D. D. Heckathorn, Probability based estimation theory for Respondent Driven Sampling, Journal of Official Statistics 24 (1) (2008) 79–97.
[12] S. Goel, M. J. Salganik, Respondent-driven sampling as markov chain monte carlo, Statistics in Medicine 28 (2009) 2202–2229.
[13] J. Illenberger, G. Flötteröd, Estimating properties from snowball sampled networks, Social NetworksAlso see www.vsp.tu-berlin.de/publications, VSP working paper number 11-01.
[14] K. J. Gile, Improved inference for respondent-driven sampling data with application to HIV prevalence estimation, Journal of the American Statistical Association 106 (493) (2011) 135–146.
[15] J.-P. Onnela, N. A. Christakis, Spreading paths in partially observed social networks, Physical Review E 85 (036106) (2012) 1–12.
[16] S. H. Lee, P.-J. Kim, H. Jeong, Statistical properties of sampled networks, Physical Review E 73 (016102) (2006) 1–7.
[17] S.-P. Wang, W.-J. Pei, Detecting unknown paths on complex networks through random walks, Physica A 388 (2009) 514–522.
[18] Q. Ye, B. Wu, B. Wang, Distance distribution and average shortest path length estimation in real-world networks, in: L. Cao, Y. Feng, J. Zhong (Eds.), Advanced Data Mining and Applications, Vol. 6440 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2010, pp. 322–333.
[19] D. Heckathorn, S. Semaan, R. Broadhead, J. Hughes, Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18–25, AIDS and Behavior 6 (2002) 55–67, 10.1023/A:1014528612685.
[20] M. J. Salganik, D. D. Heckathorn, Sampling and estimation in hidden populations using respondent-driven sampling, Sociological Methodology 34 (2004) 193–239.
[21] R. Atkinson, J. Flint, Accessing hidden and hard-to-reach populations: Snowball research strategies, Social Research Update 33.
[22] J. Johnson, J. Boster, D. Holbert, Estimating relational attributes from snowball samples through simulation, Social Networks 11 (1989) 135–158.
[23] P. Ebbes, Z. Huang, A. Rangaswamy, H. P. Thadakamalla, Sampling large-scale social networks: Insights from simulated networks, in: 18th Annual Workshop on Information Technologies and Systems, Paris, France, 2008.
[24] S. Lee, S.-H. Yook, Y. Kim, Random walks and diameter of finite scale-free networks, Physica A 387 (2008) 3033–3038.
[25] D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (440-442).
[26] M. Richardson, R. Agrawal, P. Domingos, Trust management for the semantic web, in: D. Fensel, K. Sycara, J. Mylopoulos (Eds.), The Semantic Web - ISWC 2003, Vol. 2870 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2003, pp. 351–368, 10.1007/978-3-540-39718-2_23.
[27] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Trans. Knowl. Discov. Data 1.

[28] J. Leskovec, K. J. Lang, A. Dasgupta, M. W. Mahoney, Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters, Internet Mathematics 6 (1) (2009) 29–123. arXiv:http://www.tandfonline.com/doi/pdf/10.1080/15427951.2009.10129177.