Social Sampling

Anirban Dasgupta Ravi Kumar D. Sivakumar
Yahoo! Research, 701 First Avenue, Sunnyvale, CA 94089.
{anirban,ravikumar,dsiva}@yahoo-inc.com

ABSTRACT

We investigate a class of methods that we call "social sampling," where participants in a poll respond with a summary of their friends' putative responses to the poll. Social sampling leads to a novel trade-off question: the savings in the number of samples (roughly the average degree of the network of participants) vs. the systematic bias in the poll due to the network structure.

We provide precise analyses of estimators that result from this idea. With non-uniform sampling of nodes and non-uniform weighting of neighbors' responses, we devise an ideal unbiased estimator. We show that the variance of this estimator is controlled by the second eigenvalue of the normalized Laplacian of the network (the network structure penalty) and the correlation between node degrees and the property being measured (the effective savings factor). In addition, we present a sequence of approximate estimators that are simpler or more realistic or both, and analyze their performance.

Experiments on large real-world networks show that social sampling is a powerful paradigm in obtaining accurate estimates with very few samples. At the same time, our results urge caution in interpreting recent results about "expectation vs. intent polling".

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications – Data Mining

Keywords: Polling, Social networks

1. INTRODUCTION

Polling is a method commonly employed to estimate the fraction of a population that possesses some property of interest (e.g., supporters of a political party, alcoholics, consumers of organic food, living on minimum wages, smartphone owners, gun owners, people who understand compound interest, etc.). The fundamental quantities of interest in polling are the number of samples and the error rate, and their trade-off is understood well from a statistical viewpoint: if we wish to approximate a fraction to within an additive error of ϵ, roughly $1/\epsilon^2$ samples are both necessary and sufficient.

Small subpopulations constrain the error ϵ to be small, and thus in reality the sampling cost often becomes a significant burden to deal with. In practice, researchers resort to multiple alternatives in trying to reduce this burden. For example, one could sample members of the population not uniformly but with some built-in biases that make the estimates more reliable with fewer samples; in the absence of proper normalization such schemes could often be prone to a systematic bias. In a specific instance, while estimating the fraction of schools that implement a particular program, educationalists often resort sampling schools with probability proportional to the square root of the teachers or students.*

Yet another sampling approach, and one that has found recent interest in polling research, is to inquire a respondent about how prevalent the property of interest is among the respondent's circle of friends, family, or their neighborhood, and aggregate these responses in some meaningful way. A particular example is given by economists Rothschild and Wolfers who in [14] propound the benefits of using "expectation polling" (where voters are asked about who they expect will win the election) over "intent polling" (where they are asked about their voting intent). While different such sampling methods often produce reasonable estimates in practice, few come with any kind of theoretical guarantees on either the *sampling bias* or on the *sampling error*, equivalently the number of samples required to have an ϵ-error estimate.

The framework. In this work, we investigate a class of methods we call "social sampling." The broad idea is that there is a social network underlying the population of interest, and members of the social network often have knowledge of their neighbors' predilections — social, political, as well as in other aspects of life — and will be able to summarize them meaningfully, when asked the right questions. By doing so, one could expect a savings in the number of samples required by a factor roughly equal to the d, the average degree of the social network of the participants, since a randomly chosen respondent is summarizing the results from d of his friends in order to give an answer. On the other hand, this sampling strategy now depends intimately on the network and so the analysis of the sampling bias† and error would have to incorporate the systematic effects introduced by the network structure. Our main contribution is to design the sampling strategy such that such sampling bias is overcome, and to precisely characterize the sampling error, and hence the number of samples sufficient for an ϵ-error estimate,

*See the DoE report at `http://nces.ed.gov/pubs92/92082.pdf`. In doing so, the statistical estimate one obtains is not a measure of the prevalence of the program itself, but of the number of students impacted by the program; we address this specific type of skew momentarily via volume estimates.

†This is different from the subtle bias arising from the fact that participants often tailor their notion of "friends" while answering a question, which leads to overestimation; for example, when asked how many of a respondent's friends drive a hybrid car, their notion of friendship might expand to include any acquaintance who drives a hybrid car.

KDD'12, August 12–16, 2012, Beijing, China.

in terms of both the spectral property of the network, and the correlation of the attribute and the network degree distribution.

Formally, the social sampling framework models the sampling problem as follows: $G = (V, E)$ is a social network, $f : V \to \{0, 1\}$ is a binary function that tells us whether each node in V satisfies some property, and our goal is to estimate the fraction of nodes $v \in V$ such that $f(v) = 1$. We derive several *unbiased* estimators of this fraction based on the idea of social sampling outlined above, and provide a precise analysis of the resulting estimators.

Network-induced biases. Before we describe our actual estimators in more detail, we outline the difficulties alluded to above: how does the network structure play a role in saving number of samples vs. in introducing statistical errors. Consider, for example, the network in Figure 1(a), where there is only one RED node that is connected to all the remaining $n - 1$ nodes, which are all colored BLUE. The fraction of RED nodes is vanishingly small, and yet, every node that is sampled reports having a RED node in the neighborhood; thus unless we tune the sampling method to account for network structure, we will end up with a very biased estimate.

The interplay of the *sampling error* and network structure is even more subtle. As the example in Figure1(b) shows, we might have a subpopulation of RED nodes that, although significant in overall size, is barely connected to the BLUE nodes. Furthermore, even the interconnection between the RED nodes could very well be sparse (none in Figure 1(b)). In this case, any sampling method that asks users to only proffer a summary from their neighbor's information is doomed to have a higher variance than the naive estimator, since exactly one high degree node in this graph reports a very large number of RED neighbors. Thus, the variance analysis needs to encode some characterization of the network structure, and it is not a priori obvious what could be that characterization.

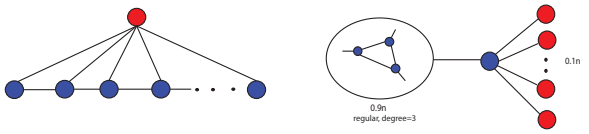

Figure 1: (a) Need to scale neighbor contributions by their degree. (b) Need for expansion: in this case neighbor querying has high variance.

Main results. We first introduce an 'ideal' unbiased estimator where nodes are sampled from the social network according a non-uniform distribution and where the function values of the nodes' neighbors are weighted in a non-uniform way (the function value at a neighbor is weighted by the reciprocal of its degree). This is intuitive since weighting the neighbors' values by the reciprocal of the degree undoes the bias introduced by large degree nodes (that will appear as neighbors of several nodes), and picking high-degree nodes as respondents of the poll naturally gives us large savings, and helps us 'reach deeper' into the network. Our main result is that the variance of this estimator is controlled by two factors: the second eigenvalue $(1 - \lambda_2)$ of the normalized Laplacian of the network, and the correlation between node degrees and the property being measured. The first factor captures the penalty to be paid because of correlation in the samples caused by the network structure, and the second factor captures the savings resulting from having a larger number of 'effective samples.' Specifically, the first factor is $m\lambda_2^2/n$, which always lies in the range $[O(1), d_{\text{avg}}]$ (d_{avg} being the average degree). The smaller λ_2 is than 1, the better connected is the network; for random regular graphs and well-connected regular expanders with degree d, λ_2^2 is roughly $1/d$, so this factor is only a constant. The second factor is $\|D^{-1/2} f\|_2^2$, where D is the diagonal matrix of the node degrees, and f is the vector representation of the function whose average is being estimated. This factor captures the correlation of f with the node degrees; again, for regular graphs, this is $\bar{f}/d$, where $\bar{f}$ is the (true) estimate and the factor d represents the savings from social sampling.

The ideal estimator assumes significant knowledge of the network, specifically about the degrees of nodes. Since this information may not be available in severely restricted contexts, we present a sequence of approximate estimators and analyze their performance (both rigorously and empirically). One simple and natural variant is when the pollster picks the nodes uniformly at random (thus eliminating the need for the knowledge of the node degrees), and the respondents incorporate their knowledge of the (putative) responses as well as the degrees of their neighbors. Another variant of interest is one where the pollster (e.g., a social networking site) has perfect knowledge of the network structure (except for the function values), but it is unrealistic to expect nodes to have perfect information about their neighbors' degrees. In this case, the pollster picks nodes according to the ideal estimator, but for each chosen node, samples a small number of its neighbors, and asks the chosen node a query of the form "among your friends A, B, and C, how many exercise regularly?" In both of these highlighted cases, we derive unbiased estimators and analyze their variances.

A third variant attempts to formally model "expectation polling," a subject of interest in polling research [14]. The authors presented empirical analysis to conclude that expectation polling is often significantly more accurate than intent polling. As an explanation, they suggested that when polled about their expectations, respondents implicitly summarize the intents of their friends and family. Our analysis shows that if this is carried out on a social network, the result will incorporate each person's intent with a weight proportional to its degree in the social network; thus, well-connected nodes might end up skewing the poll result significantly. However, it is incorrect to assume that in an expectation poll, the respondents include all of their acquaintances with equal importance; it is natural for a respondent to "discount" one of their neighbors if the neighbor is an exceptionally popular person (unless they are truly close to the respondent). We model this in the following probabilistic way. When polled, each person implicitly creates an "effective set" of their friends to which each of their neighbors belongs with probability inversely proportional to its (the neighbor's) degree. The polled person then tosses a coin with probability of heads equal to the fraction of this effective set that is RED, and reports either RED or BLUE depending on the outcome, and also announces the size of the effective set. We show that this model, too, enjoys the savings in the number of samples similar to the ideal estimator; our analysis of this model complements [14] and offers a way to rigorously analyze the success of expectation polls.

Another variant of the estimators touches upon the difference between two useful ways to reason about the popularity of a concept within a society. One way, of course, is the *prevalence* of the concept, measured as the fraction of the set of members to whom the concept applies; another way is to measure the *pervasiveness* of the concept, measured as the fraction of the pairwise relations such that at least one of members involved subscribes to the concept. Often, the prevalence and pervasiveness of a concept are wildly different, depending on how influential its adopters are. We show that pervasiveness can be modeled by the *volume* of the set of nodes[‡] with the given property, and show how to estimate this quantity efficiently (by a much simpler variant of the ideal estimator).

[‡]Formally, volume of a set $S \subseteq V$ of nodes in a network is the total degree of the nodes in S.

Finally, we apply the estimators to several large real-world graphs, and report experimental results. One of the networks is a snapshot of around 4M nodes from the LiveJournal social network, where we estimate the prevalence of various natural properties — users from geographic locations of various magnitudes (cities, states, countries), users of various age groups, users interested in a variety of topics, etc. The other network we study is the coauthorship network among roughly 1M authors whose papers are included in the DBLP database; here we estimate the prevalence of various publishing outlets (conferences and journals). Our experiments show that social sampling is a powerful paradigm in obtaining accurate estimates from significantly fewer samples than standard sampling. At the same time, they also highlight how the performance of the estimators are sensitive to the structure of the network, and how prevalent and pervasive the properties are.

2. RELATED WORK

Sampling to estimate hidden subpopulation is classical. The fact that an ϵ-additive error can be achieved with constant probability using $O(\frac{1}{\epsilon^2})$ samples can be shown using a variety of tail inequalities and is folklore. Lower bounds on the number of samples, on the other hand, is more recent [4, 2]. Sampling algorithms have been studied in large networks, from a perspective of estimating different network properties. Leskovec and Faloutsos [12] defined the problem of creating a representative sample of the network that approximates multiple properties of the original network such as average shortest path, centrality, etc; see also [5, 13]. An intriguing variant of the network sampling question was posed by Backstrom and Kleinberg [1] in the context of creating buckets of social network users for A/B testing of features that require social interactions. There is an implicit trade-off of independence in sampling and correlation (due to ties to chosen samples) that they study; see also [11]. Voting on social networks has been considered by Boldi et al. [3] who studied different vote delegation mechanisms.

Katzir, Liberty, and Somekh [10] presented a method to estimate the size of online social networks using degree-biased sampling and counting collisions in the samples collected; their method can also be used to estimate subpopulations. Their technique is more geared for settings where the sampling is restricted to using the public API provided by the social networks; for instance, choosing an uniformly random node is expensive in their model since it requires rejection sampling.

In the theoretical domain, Goldreich's survey [7, 8] provides a nice elucidation of the trade-off between the number of random bits utilized, the number of samples, and the computational complexity in sampling to estimate a function within an ϵ-additive error. In fact, our guarantees in Theorem 4 and Theorem 12 are generalizations of a result [8] on regular (large degree) expanders, and show the same order of improvement in sample size. Our model, though is different, in that we are not interested in saving randomness and we consider neighbor queries by sampled nodes to be free.

As mentioned before, the question inspiring our work is closest to the one studied by Rothschild and Wolfers [14], where they study the value of asking "who is expected to win" as opposed to "whom do you expect to vote for" and show empirical evidence that the former often leads to better estimates; they do not investigate the question in the context of explicit social networks or present theoretical quantification of when such a gain could be expected.

The presence of social connections among the sampled population has also been exploited in the *snowball sampling* class of techniques of which *respondent driven sampling* (RDS) [9] is probably the best known. RDS has proven to be useful in practice, having being used in over 120 studies. RDS, however, has not yet been amenable to theoretical analysis, except under assumptions of random graph models [6]. Our technique is orthogonal to RDS (and in principle composable) since the purpose there is to use the network to recruit the next respondent, not reduce the number of samples by using people's knowledge of their neighbors' predilections.

There have also been numerous studies on the interaction of network structure and node attributes. One theory that deserves a mention in our setting is that of *homophily* or *assortativity*: the tendency of nodes that are alike to link to each other. The network structure in fact has a much more subtle effect on the sampling error than a simple homophily metric can measure. The effect of increasing homophily on our estimates is obtained through the second eigenvalue characterization: if increasing homophily makes the network more partitioned, our estimates will degrade. In Section 4, we show some experiments with attributes that have different homophilies.

3. ALGORITHMS

3.1 Notation

The network is represented by an undirected graph $G = (V, E)$, where V denotes the set of nodes and E denotes the set of edges. Let $n = |V|$ and $m = |E|$. Frequently, the variables u, v will denote the nodes of G. For any node u, the set of neighbors of u is denoted by $N(u)$. The *degree* of node u is denoted $d_u = |N(u)|$. We will assume that $d_u \geq 1$ for all nodes. Let $f : V \to \{0, 1\}$ denote the function of interest, i.e., we wish to estimate the fraction $\bar{f} = \frac{1}{n}|\{u \mid f(u) = 1\}|$. We will also consider f as the indicator (column) vector of the set $\{u \mid f(u) = 1\}$, i.e., $f_u = f(u)$. Let A denote the adjacency matrix of the graph, i.e., $A_{uv} = 1$ if and only if $(u, v) \in E$. Let D be a diagonal matrix such that $D_{uu} = d_u$. For any vector $p \in \mathbb{R}^n$ such that $\sum_u p_u = 1$, we define the diagonal matrix P by $P_{uu} = p_u$. Finally, let $\mathbb{1}$ denote the n-element vector of all 1's, i.e., $\mathbb{1}_u = 1$ for all u.

3.2 Estimating size

The most basic algorithm for estimating $|\{u \mid f(u) = 1\}|$, or equivalently the fraction $\bar{f}$, is the standard estimator (Algorithm 1) that samples a set of nodes and polls each sampled node u to check if $f(u) = 1$, and reports the fraction of the nodes that satisfy this. The following guarantee on the Naive estimator is well-known [7].

Algorithm 1 Naive size estimator.

Require: Graph G, function $f : V \to \{0, 1\}$, sample size r.
Choose a random sample $S \subseteq V$ of r nodes with replacement by including node u with probability $\frac{1}{n}$.
Return $\hat{f} = \frac{1}{r} \sum_{u \in S} f(u)$.

FACT 1. *If we pick $r = 2/(\epsilon^2 \delta)$ samples, with probability $1-\delta$, the* Naive *estimator will give an estimate $\hat{f}$ such that $|\hat{f} - \bar{f}| < \epsilon$.*

The intuition behind "social sampling" is that if we poll a node u to obtain some estimate of $\{f(v) \mid v \in N(u)\}$, we could use fewer samples than we need with the naive estimator. As the example in Figure 1 showed, we need to normalize by the degree of the node in order to get an unbiased estimator. The next sampler we present, the ideal size estimator (Algorithm 2), corrects precisely for this bias, by scaling $f(v)$ by $1/d_v$.

We now proceed to analyze the quality of the ideal estimator for different distributions p. We do this by establishing a lemma that bounds the variance of the above estimator; first, of course, we note that the ideal estimator is an unbiased estimator of $\bar{f}$.

Algorithm 2 Ideal size estimator(p).

Require: Graph G, function $f : V \to \{0,1\}$, probability distribution p on V, and sample size r.
Choose a random sample $S \subseteq V$ of r nodes with replacement by including node u with probability p_u.
Return $\hat{f} = \frac{1}{nr} \sum_{u \in S} \frac{1}{p_u} \sum_{v \in N(u)} f(v)/d_v$.

Define $e_u \in \mathbb{R}^n$ to be the vector with only $e_u(u) = 1$ and $e_u(v) = 0$ for $v \neq u$. Define the random variable F as follows: pick $u \in V$ with probability p_u, and let $F = \frac{e_u^T A D^{-1} f}{n p_u}$, which is well defined since by assumption $d_v \geq 1$ for all v. To help the analysis, note that the vth component of the vector $D^{-1}f$ is precisely $f(v)/d_v$, and the uth component of the vector $AD^{-1}f$ is precisely $\sum_{v \in N(u)} f(v)/d_v$. Recall that P is a diagonal matrix such that $P_{uu} = p_u$. The following establishes the unbiasedness property, independent of p, and characterizes the variance in terms of a quantity that we will later bound.

LEMMA 2. *The random variable F satisfies $\mathrm{E}[F] = \bar{f}$ and $\mathrm{E}[F^2] = \frac{1}{n^2} f^T D^{-1} A P^{-1} A D^{-1} f$.*

PROOF. The expectation can be simply computed as

$$\mathrm{E}[F] = \sum_u p_u \frac{e_u^T A D^{-1} f}{n p_u} = \sum_u e_u^T A D^{-1} f / n$$
$$= \mathbb{1}^T A D^{-1} f / n = D D^{-1} f / n = \bar{f}.$$

For $\mathrm{E}[F^2]$, we have

$$n^2 \mathrm{E}[F^2] = \sum_u p_u \left(\frac{e_u^T A D^{-1} f}{p_u} \right)^2$$
$$= \sum_u \frac{(e_u^T A D^{-1} f)^2}{p_u}$$
$$= \sum_u f^T D^{-1} A e_u e_u^T A D^{-1} f / p_u$$
$$= f^T D^{-1} A \left(\sum_u e_u e_u^T / p_u \right) A D^{-1} f = f^T D^{-1} A P^{-1} A D^{-1} f. \quad \square$$

Next, we bound its variance for the special case when respondents are sampled with probability proportional to their degree.

LEMMA 3. *Let the sampling probability $p_u = \frac{d_u}{2m}$. Define the matrix L by $L = D^{-1/2} A D^{-1/2}$ and let λ_2 denote the second largest eigenvalue of L.[§] Then $\mathrm{var}(F) \leq (2m/n^2)\lambda_2^2 \|D^{-1/2} f\|^2$.*

PROOF. When the sampling probability p_u satisfies $p_u = d_u/2m$, we have $P = D/2m$.

$$\mathrm{var}(F) = f^T D^{-1} A P^{-1} A D^{-1} f / n^2 - \bar{f}^2$$
$$= f^T D^{-1} A (2m D^{-1}) A D^{-1} f / n^2 - \bar{f}^2$$
$$= \frac{2m}{n^2} f^T D^{-1} A D^{-1} A D^{-1} f - \bar{f}^2$$
$$= \frac{2m}{n^2} f^T D^{-1/2} L^2 D^{-1/2} f - f^T \mathbb{1}\mathbb{1}^T f / n^2$$
$$= \frac{2m}{n^2} f^T D^{-1/2} \left(L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m} \right) D^{-1/2} f$$
$$\leq \frac{2m}{n^2} \left\| L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m} \right\| \left\| D^{-1/2} f \right\|^2.$$

[§] $I - L$ is called the *normalized Laplacian* and thus $1 - \lambda_2$ is the second smallest eigenvalue of the normalized Laplacian.

Now, it can be verified by basic manipulations that the first eigenvector of L^2 is $\frac{D^{1/2}\mathbb{1}}{\sqrt{2m}}$ with eigenvalue 1. Thus we have $\|L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m}\| = \lambda_2(L^2) = \lambda_2^2$. The proof follows. $\square$

Using this variance bound, we can easily obtain a bound on the number of samples sufficient in order to get an ϵ-error estimate.

THEOREM 4. *With $p_u = \frac{d_u}{2m}$ and sample size $r = \lceil \frac{4m\lambda_2^2 \|D^{-1/2} f\|^2}{n^2 \epsilon^2 \delta} \rceil$, the* Ideal *size estimator produces $\hat{f}$ such that $|\bar{f} - \hat{f}| < \epsilon$ with probability $1 - \delta$.*

PROOF. If we take r samples and denote the corresponding estimators by $F_1, \ldots F_r$, then, $\hat{f} = \sum_i F_i / r$. Applying the Chebyshev inequality, we have

$$\Pr[|\hat{f} - \bar{f}| > \epsilon] \leq \frac{2\mathrm{var}(\hat{f})}{\epsilon^2} \leq \frac{2\mathrm{var}(F)}{r\epsilon^2}.$$

Choosing $r = \frac{2\mathrm{var}(F)}{\epsilon^2 \delta} = \frac{4m\lambda_2^2 \|D^{-1/2} f\|^2}{n^2 \epsilon^2 \delta}$, the proof follows. $\square$

If the initial sampling probabilities are changed from being proportional to degree to being proportional to the square roots of degrees, we obtain a similar bound. Instead of presenting a very similar proof again, we summarize the result below.

COROLLARY 5. *Let $m_s = \sum_u \sqrt{d_u}$. If we take $p_u = \frac{\sqrt{d_u}}{m_s}$ and $r = \lceil d_{\max}^{1/2} \frac{2m_s \lambda_2^2 \|D^{-1/2} f\|^2}{n^2 \epsilon^2 \delta} \rceil$ samples, the* Ideal *estimator satisfies $|\hat{f} - \bar{f}| < \epsilon$ with probability $1 - \delta$.*

The number of samples required for uniform sampling can be bound similarly as follows. Its proof follows from the proofs of Corollary 5 and Theorem 4, by plugging in the suitable probabilities.

COROLLARY 6. *For sampling probabilities $p_u = \frac{1}{n}$, the variance of the sampler can be bounded by $\mathrm{var}(F) \leq \frac{1}{n} f^T D^{-1} A^2 D^{-1} f - \bar{f}^2$. In order to get an ϵ-error estimate with probability $1 - \delta$, it suffices to take $r = \lceil d_{\max} \frac{2\lambda_2^2 \|D^{-1/2} f\|^2}{n\epsilon^2 \delta} \rceil$ samples.*

The actual gain obtained by using the degree-based sampler instead of the Naive estimator can be clearly seen if the graph is regular, and more specifically, an expander. Recall that $\lambda_2 < 1$ always.

COROLLARY 7. *Suppose G is a d-regular graph. Then the* Ideal *estimator with probability distribution $p_u = \frac{1}{n}$ requires $\lceil \frac{\lambda_2^2}{\epsilon^2 \delta} \rceil$ samples to produce an ϵ-error estimate of $\bar{f}$ with probability $1 - \delta$. If in addition G is also an expander, with $\lambda_2 = cd^{-1/2}$, then the* Ideal *estimator requires $\lceil \frac{c^2}{d\epsilon^2 \delta} \rceil$ samples.*

From the preceding analyses, it follows that the correlation between the function f on V and the degrees of the nodes in V is one source of variance. We consider an interesting special case, where there is no such correlation, i.e., when f is obtained by tossing independent coins with probability $\bar{f}$ for each $u \in V$. In this case, we can actually compute the optimal strategy for sampling the nodes. It is obtained by minimizing the resulting variance.

THEOREM 8. *If for each node u $f_u = 1$ as a result of a i.i.d. coin toss with probability $\bar{f}$, then the optimal strategy to include node u in the sample with probability p_u proportional $\sum_{v \in N(u)} \frac{1}{d_v}$.*

PROOF. Under a sampling strategy given by distribution p, the variance is given by

$$\mathrm{var}_p(F) = \sum_u \frac{1}{p_u} \left(e_u^T A D^{-1} f \right)^2 - \bar{f}^2.$$

Thus

$$\mathrm{E}_f[\mathrm{var}_p(F)] + \bar{f}^2 = \sum_u \frac{1}{p_u}\mathrm{E}_f[(e_u^T AD^{-1}f)^2]$$

$$= \sum_u \frac{1}{p_u}\mathrm{E}_f\left[\sum_{v\in N(u)} f_v/d_v\right]^2 = \sum_u \frac{1}{p_u}\mathrm{E}_f\left[\sum_{v,v'\in N(u)} f_v f_{v'}/d_v d_{v'}\right]$$

$$= \sum_u \frac{1}{p_u}\sum_{v,v'\in N(u)} \bar{f}^2/d_v d_{v'} = \bar{f}^2 \sum_u \frac{1}{p_u}\left(\sum_{v\in N(u)} \frac{1}{d_v}\right)^2.$$

Thus, we need to find the optimal set of probabilities p_u such that the above expression is minimized under the constraint $\sum_u p_u = 1$. By using a Lagrangian it is possible to show that $p_u \propto \sum_{v\in N(u)} \frac{1}{d_v}$ minimizes the above expression. □

3.3 Approximating the Ideal estimator

The Ideal estimator and its variants incorporate the idea that a sampled node u has enough knowledge about its neighbors v to give the pollster responses based on $f(v)$ and d_v (for all its neighbors v). Next we introduce a variant that we motivate in two ways. The common outcome in both cases is that each node u only includes a small number of its neighbors in producing its response to the poll. The first setting is where the pollster has complete knowledge of the network (e.g., an online social network like Facebook), including the neighborhood structure and the degrees of all nodes. The only unknown is f. What is expensive in this model is asking a node u about too many of its neighbors; it would be desirable to poll a node u about $f(v)$ for a small number (say five) of carefully chosen neighbors v of u.¶ The second motivation for this style of sampling is that it serves as a model for polls of the form "Think of three of your best friends, and then tell us how many of them are left-handed." The hypothesis is that when a person is asked a question of this form, they will tend not to pick their highly popular neighbors, so the distribution that picks neighbors with probability inversely proportional to their degrees may be thought of as a good model for the distribution with which they pick their friends.

Algorithm 3 Sparse size estimator(p).

Require: Graph G, function $f : V \to \{0,1\}$, probability distribution p on V, and sample size r.
Choose a random sample S of r nodes with replacement by including node u with probability p_u.
For each node $u \in S$, create a set $T_u \subseteq N(u)$ by picking each neighbor $v \in N(u)$ with probability $1/d_v$.
For each node $u \in S$, create a set $T'_u \subseteq T_u$ by picking k of its neighbors from T_u (without replacement).
For $u \in S$, compute $H_u = (|T_u|/k)\sum_{v\in T'_u} f(v)$.
Return $\hat{f} = \frac{1}{nr}\sum_{u\in S}\frac{1}{p_u}H_u$.

Deriving a variance bound on the error of this sampler is harder. We just claim that this is an unbiased estimator. The proof is a simple application of the linearity of expectation.

LEMMA 9. *Algorithm 3 obtains an unbiased estimate of $\bar{f}$.*

3.4 A model of expectation polling

Next we present a model of "expectation" polling, a type of polling used for predicting elections, where the participants are asked not about their voting *intent*, but about who they *expect* will win the election. Rothschild and Wolfers [14] have recently argued that expectation polling is more powerful than intent polling, and have suggested that the power of expectation polling comes from the fact that the respondents often give an aggregate view of their friends and family (i.e., neighbors in a social network).

We propose a model of how respondents in expectation polls summarize the intent of their neighbors in a social network (Algorithm 4). The idea is that the polled node, instead of giving just a 0/1 vote, gives a weighted vote: either zero or an estimate of the neighborhood belonging to $f = 1$. We can show a similar bound to Theorem 4, using a similar variance computation.

Algorithm 4 Expec: winner estimator(p).

Require: Graph A, function $f : V \to \{0,1\}$. Probability of sampling node u is p_u and size of sample is r.
Choose a random sample S of r nodes with replacement by including node u with probability p_u.
Node u when picked, computes $q_u = \sum_{v\in N_u}\frac{f_v}{d_v}$ and $r_u = \sum_{v\in N_u}\frac{1}{d_v}$ and returns G_u where $G_u = r_u$ with probability $\frac{q_u}{r_u}$ and 0 else.
Return $\hat{f} = \frac{1}{nr}\sum_{u\in S}\frac{1}{p_u}G_u$.

LEMMA 10. *For $p_u = \frac{d_u}{2m}$, if we use $\lceil \frac{4m\lambda_2^2\|D^{-1/2}f\|\|D^{-1/2}(\mathbb{1}-f)\|}{n^2\epsilon^2\delta}\rceil$ samples, the resulting estimate $\hat{f}$ satisfies $|\bar{f} - \hat{f}| < \epsilon$ with probability $1-\delta$.*

3.5 Estimating volume

The fraction of nodes that satisfy some property $f : V \to \{0,1\}$ is an estimate of the prevalence of a property within the nodes of a social network. However, the participants modeled as nodes in a social network interact with each other, and often have knowledge of their neighbors. This motivates the problem of measuring the *pervasiveness* of a property, measured not just as the fraction of nodes that possess that property, but taking into account all the pairwise interactions in which at least one node has $f = 1$. The intuition here is that this represents the amount of "mindshare" this property has in the social network, assuming interactions happen only over, and uniformly across, the existing edges. To capture this, we invoke the concept of the *volume* of the function $f : V \to \{0,1\}$, defined by $d_f = \sum_{u:f(u)=1} d_u$; in order to estimate volume, we use the same set of analysis and methods. The main difference is that now each sampled node u only needs to return $\sum_{v\in N(u)} f(v)$ instead of the degree-normalized sum as in Algorithm 2. The lemma corresponding to Lemma 2 is the following. Define again the random variable G as $G_u = e_u^T Af/2mp_u$ with probability p_u.

Algorithm 5 Ideal volume estimator(p).

Require: Graph A, function $f : V \to \{0,1\}$. Probability of sampling node u is p_u and size of sample is r.
Choose a random sample S of r nodes with replacement by including node u with probability p_u.
Return $\widehat{d_f} = \frac{1}{2mr}\sum_{u\in S}\frac{1}{p_u}\sum_{v\in N(u)} f(v)$.

LEMMA 11. *For the random variable G, $\mathrm{E}[G] = d_f/2m$, and $\mathrm{var}[G] = (1/4m^2)(f^T AP^{-1}Af - d_f^2)$.*

The proof follows by calculation similar to Lemma 2. And thus, we obtain the following result for degree-based sampling.

¶In fact, the pollster only needs to ask how many of the five chosen neighbors have $f(v) = 1$.

THEOREM 12. *With $p_u = d_u/2m$, if we use $\lceil \lambda_2^2/(\epsilon^2\delta) \rceil$ samples, then the resulting estimate $\widehat{d}_f$ satisfies $|\frac{d_f}{2m} - \frac{\widehat{d}_f}{2m}| \leq \epsilon$ with probability at least $1-\delta$.*

PROOF. (Sketch) First, choosing $p_u = d_u/2m$, the variance of the estimator is

$$\begin{aligned}
\mathrm{var}(G) &= \frac{1}{4m^2} f^T A P^{-1} A f - \frac{d_f^2}{4m^2} \\
&= \frac{1}{2m} f^T A D^{-1} A f - \frac{D\mathbb{1}\mathbb{1}^T D}{4m^2} \\
&= \frac{1}{2m} f^T D^{1/2} \left(D^{-1/2} A D^{-1} A D^{-1/2} f - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m} \right) D^{1/2} f \\
&= \frac{1}{2m} f^T D^{1/2} \left(L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m} \right) D^{1/2} f \\
&\leq \frac{1}{2m} \|f^T D^{1/2}\|^2 \|L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m}\| \text{ using Cauchy–Schwarz.}
\end{aligned}$$

Note that $\|f^T D^{1/2}\|^2 = d_f$ and that $\|L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m}\|$ equals $\lambda_2(L^2) = \lambda_2^2$. Thus $\mathrm{var}(G) \leq \frac{1}{2m} d_f \lambda_2^2$. Recall $\widehat{d}(f) = \sum_i G_i$ where G_i is an independent copy of G. Using the Chebyshev inequality, we have

$$\Pr\left[\left| \frac{\widehat{d}_f}{2m} - \frac{d_f}{2m} \right| > \epsilon \right] \leq \frac{\mathrm{var}(G)}{r\epsilon^2} \leq \frac{d_f \lambda_2^2}{2mr\epsilon^2}.$$

Since $d_f \leq 2m$, choosing $r = \frac{\lambda_2^2}{\epsilon^2 \delta}$, we have the result. □

Similarly, when f is the result of a random coin toss, the optimal set of sampling probabilities are now obtained by sampling each node proportional to its degree.

THEOREM 13. *If for each node u, $f_u = 1$ as a result of an i.i.d. coin toss with probability $\bar{f}$, then the optimal strategy to sample in order to estimate d_f is to sample node u with probability proportional to d_v.*

Again, the proof works by writing down the expression for the $\mathrm{var}(F_p)$ from Algorithm 5 and choosing the appropriate probability distribution to minimize this quantity for a random f.

4. EXPERIMENTAL SETUP

Data. Our dataset consists of large networks where each node in the network is associated with one or more attributes. Specifically, we consider the following two networks.

The LIVEJOURNAL network is built from the social network LiveJournal (`livejournal.com`). The network consists of about 5.36M nodes and about 160M directed edges and is a version of the social network crawled in March 2008 (the entire data is available through a public API). For our purposes, we treat the network as an undirected network by dropping the edge directions. In addition to the network, we also extract the following attributes (if available) for each of the users from their public profiles: age, location, and a list of interests expressed as free text (e.g., cricket, tennis). The locations were parsed using Yahoo! geolocation API to filter out invalid ones and extract the city, state, country information. Note that not all the nodes have all the attributes. About 58% of the users have a valid location, 40% of the users of have a valid age, and 60% of the users have at least one interest.

The DBLP network is built from the DBLP database (`http://www.informatik.uni-trier.de/~ley/db/`). This data is also available publicly. The network consists of 968K nodes and about 8M undirected edges, where the nodes correspond to authors and the edges corresponding to coauthorship on some publication. We extract all the publication venues of an author and these (free text, abbreviations) form the node attributes. By construction, every node has at least one attribute and there are over a million distinct attributes. We sampled 100 of these attributes with probability proportional to their popularity and used these in our experiments.

Error. In all our experiments, we measure the absolute error of our estimator with respect to the true value, i.e., if the output of our estimator is $\tilde{z}$ and the true value is z, we measure the error as $|z - \tilde{z}|$; in all our experiments, $z \in [0,1]$. Recall that our estimators are designed to output an additive $\pm\epsilon$-approximation to the truth and hence measuring the absolute error is valid. Estimators that work with relative accuracy will require a lot more samples: to measure the bias p of a coin to within $(1 \pm \epsilon)$ error requires $\Omega(\frac{1}{p\epsilon^2})$ samples.

Sampling. We chose a set of number of samples, from a minimum of 100 samples to a maximum of 50000 samples. For a specified number k of samples, we will sample k nodes from the network according to various distributions. We consider the following distributions: uniform at random (unif), proportional to the degree of the node (deg), proportional to the square root of the degree of the node (sqrtdeg), and proportional to the reciprocal of the sum of the degree of the neighbors of the node (recdeg).

Given a set A of k samples and an estimator g, we proceed in the following way. Let $L = \{1, 3, 5, 7, 9\}$. For each $\ell \in L$, we first chop A into ℓ partitions $A_1, \ldots, A_\ell$, apply the estimator $g(\cdot)$ on each of the partitions, and then output the median of the results, i.e., we output $g'_\ell(A) = \mathrm{med}\{g(A_1), \ldots, g(A_\ell)\}$. We then output both $\min_{\ell \in L} g'_\ell(A)$ as well as $\mathrm{mean}_{\ell \in L} g'_\ell(A)$. For simplicity, we report all our results with respect to the latter; in Section 5.6, we show that using the min vs mean does not affect the results/trends.

Estimators. We called the basic estimator Naive: recall that this estimator samples nodes uniformly at random (unif) and outputs an estimate based on how many of the sampled nodes have the attribute. We refer to the ideal estimator as Ideal, the estimator that is meant to capture the social expectation of a user as Expec, and the estimator based on a sampling of close friends as Sparse.

5. RESULTS

We summarize the results of our experiments on the two datasets. Recall than in LIVEJOURNAL, the attributes of interest were the geo (each of five cities, states, and countries), a set of ages, a set of age-buckets and a set of interests (various sports). In each of the cases, the typical size of the attribute set ranged from 0.4% to 4%, with the maximum country size being 38% and the maximum age-bucket size being 30% of the entire network. In each of the cases, the additive errors plotted are typically smaller than the average size of the attribute in that category. In each of the cases, in order to generate the curves, after computing the grouping described above, we plot the mean error over multiple groupings. To simplify presentation, we aggregated the results of the geo attributes, those of the age attributes, and the age-bucket attributes.

In DBLP, the set of attributes corresponded to venues of publication, and the set of 100 selected ranged in size from 0.01% to 3% in size, with a median of 0.5%. Again, each of our curves is obtained by averaging over all these attributes.

To clarify the trend better, we also smoothed the curves using Bezier smoothing in `gnuplot`; this lets us compare the trends without being too concerned about the variance issues that are bothersome at this small scale. For larger scale experiments, the corresponding sample sizes would be larger and variance would be less.

5.1 Performance on LIVEJOURNAL

The results for the LIVEJOURNAL dataset for the geo attributes are summarized in Figure 2. They show the performance of the sampling algorithms on the city, state, and country attributes. The algorithms considered are the Naive estimator with uniform bias, the Ideal estimator with degree and square-root degree biases, the Sparse and Expec both with degree biases. Any other biases would only increase the variance of the Naive estimator. The performance of the estimator that has sampling bias of sum of inverse-degrees of neighbors is almost always subsumed by the degree-biased and square-root biased estimators, and furthermore, is computationally more expensive. Thus, we leave out this inverse-degree-sum estimator in most experiments except for one in Figure 5.

The x-axis is the sample size and the y-axis is the additive error. From the plot, it is clear that the performance of Ideal for the degree and square-root-degree biases are the same. The performance curves all exhibit similar trends; as expected, the error decreases faster than linearly with an increase in the number of samples.

For the city attributes, the performance of Naive is comparable to that of Expec and is significantly worse than that of Ideal. In the sample size range up to 10k, the error incurred by Naive is about 200% that of the Ideal sampler. In terms of the absolute error size, this corresponds to about 0.1% of the network even near sample size 1k. For the smaller attributes that we have, this difference is essentially one between detecting or not detecting its presence at all. If we fix the error, Naive requires at least twice as many or more samples to achieve the same error as Ideal does. Note that at 10k samples, we are already at 2% of the entire network, a significant sample size. The performance of Sparse (with sparsity parameter $= 5$) is similar to that of Naive. Also, Expec seems to have a non-monotone behavior near sample size 200, that is again caused by the variance due to small sample sizes: for a fixed sample size Expec has a higher variance than the other estimators due to the discretization of the value returned by each node.

For the state and country attributes, the performance of the samplers are very much similar: Naive performs comparably with degree and square-root biases, and these three perform slightly better than both Expec and Sparse. For states, square-root performs marginally better than every one else, and performs better than degree in both. This is interesting as the theoretical bound we have for the variance of square-root bias is strictly worse than the theoretical bound for the degree bias. Different factors could be contributing to the disparity of the theoretical and empirical performance: the variance bound given is only a crude bound that does not capture the real interaction of f with the structure of G (upper bounds the interaction using a heavy-handed Cauchy–Schwarz inequality). Furthermore, even though the bound given is in terms of λ_2, this quantity itself is not small for these networks since these networks are not known to be expanders at the small scales ($\lesssim$ 10k), which in fact are the attribute sizes that we mostly have.

Figure 3 summarizes the results for the age- and interest-related attributes. Five age attributes (20, 25, 30, 35, and 40) and five age buckets $([20, 29), \ldots,)$ were considered. Once again, the performance of Ideal with degree and square-root degree are much better than that of Naive, offering the same error bound at about half the number of samples. Expec performs worse than Naive in these cases. Together with the results of Figure 2, the indifferent performance of Expec suggests that our model of expectation polling leads to high-variance estimators; to the extent that our model reflects the reality of expectation polling, this is a caveat in interpreting the recent excitement about expectation polling. The final set of attributes for LIVEJOURNAL are a set of sports-related interests, tennis, cricket, etc. Here the performance of the estimators converge sooner, and after sample size 2k, they are virtually identical.

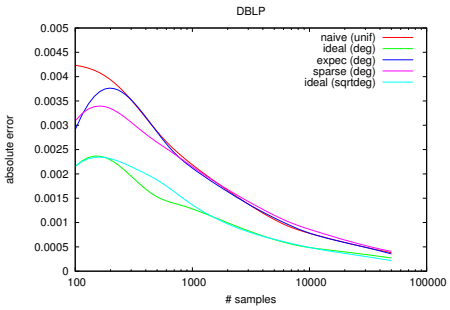

Figure 4: Performance of various algorithms on DBLP.

5.2 Performance on DBLP

For the DBLP network, the attributes refer to publication venues, as described above. We randomly selected 100 such attributes from the set of million attributes with probability proportional to their popularity. The parameter settings for grouping to compute the median are similar as in Section 5.1.

In the DBLP network, the improvement offered by the Ideal set of samplers is starker. For the sample size, the error by the Naive sampler is again 200-300% of the error by the ideal sampler, and Naive requires at least 3-5 times the samples that Ideal uses to achieve the same error. In fact, for the DBLP network, the performance samplers do not converge even at 20k samples, which is about 4% of the entire set of nodes, a pretty big sample size.

5.3 Effect of the initial node selection bias

We now investigate, in Figure 5, the effect of the initial node selection bias in more details, in the context of the cities attribute in LIVEJOURNAL, and for two sampling strategies, Ideal and Expec. We first consider the Ideal algorithm and investigate the different node selection biases — uniform, degree, square-root of degree, and sum of reciprocal-degrees of neighborhood. It is interesting to see that the performances quickly converge, even at sample size 500, with the reciprocal-sum degree and the square-root performing only marginally than the others till sample size 10k. For the Expec sampler, we observe the same phenomenon, but here the uniform strategy is slightly worse, and the reciprocal-sum slightly better than the all the rest.

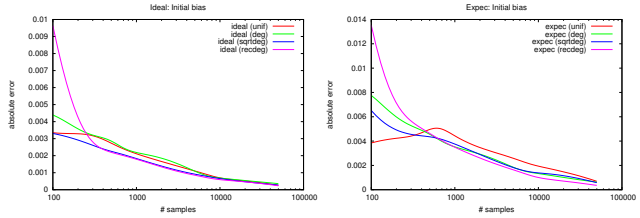

Figure 5: Performance of Ideal and Expec algorithms on the LIVEJOURNAL data for different initial node selection bias.

5.4 Homophily vs. Homogeneity

A natural question to ask is whether the sampling error is correlated with the homophily of the attribute f. In this section, we show that a better way of characterizing the error is not in terms of

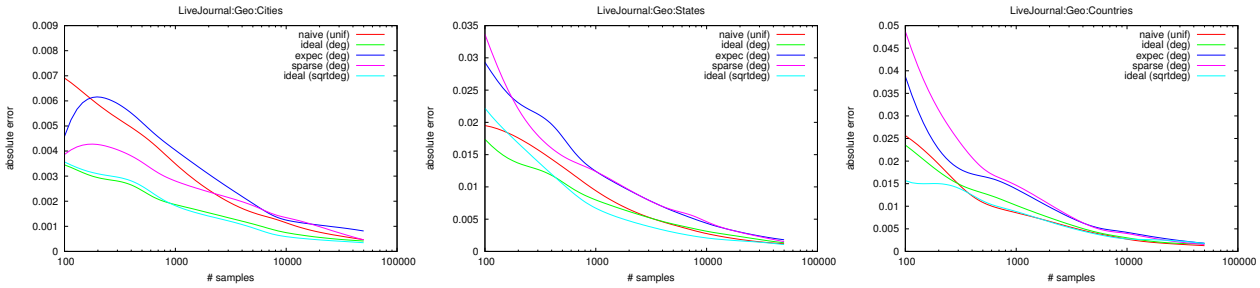

Figure 2: Performance of various algorithms on the LIVEJOURNAL data for random cities, states, and countries.

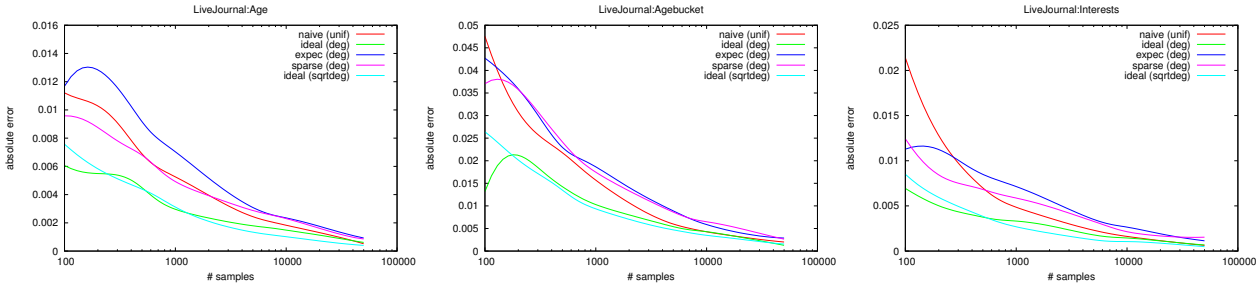

Figure 3: Performance of various algorithms on the LIVEJOURNAL data for random ages, age buckets, and interests.

	f	Naive	Ideal
spearman	0.223	0.126	0.212
pearson	0.086	0.087	0.121

Table 1: Correlation of CH with attribute size, Naive and Ideal errors using both rank (spearman) and a moment (pearson) methods.

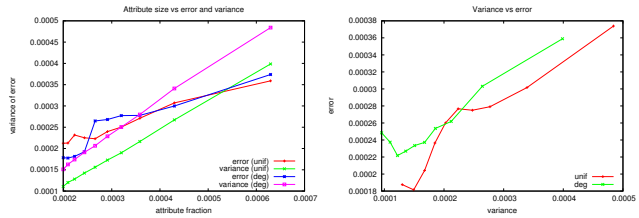

Figure 6: Relation of error and variance for uniform and degree-biased selection and for different attribute sizes.

homophily, but rather in terms of a weighted measure of "homogeneity" of neighbors — whether neighbors of a node are of same color. In investigating the effect of homophily on error, we measure homophily by the following metric, commonly known as the *Coleman homophily index*.

$$H_f = \frac{\sum_{u,v \in f} 2A_{uv}}{\sum_{u,v \in f} 2A_{uv} + \sum_{u \in f, v \in V \setminus f} A_{uv}}, \quad \mathrm{CH}_f = \frac{H_f - \bar{f}}{1 - \bar{f}}.$$

Note that $\mathrm{CH}_f \in [0,1]$ and is zero if f is the result of a random coin toss for every node. Table 1 shows the correlation between the homophily index of 241 cities with the size of the city and with the Naive and Ideal errors. If there was a homophily effect independent of size, then we would see the correlation between Ideal and CH to be higher than the correlation of CH with both Naive-error and $\bar{f}$. From Table 1, the effect of attribute homophily on the sampling error does not seem to be significant.

The sampling error, on, the other hand, depends on the variance of the estimator, as expressed in Lemma 2. In Figure 6 we investigate relation of the variance and the actual error using the same 241 cities. For the variance bound, we use the expression in Lemma 2, that is, $\mathrm{E}[F^2] - \bar{f}^2$. In order to generate this plot, we grouped the 241 attributes into 10 buckets based on the variance quantiles — then the average in each bucket error was plotted against the average variance in each bucket. This was done for both the uniform and the degree bias. In both case, the error has mostly a linear relation with the variance, showing that to a reasonable degree of approximation, bounding the variance is the right strategy in order to minimize the error. Furthermore, the network structure interpretation that the variance encodes is one that is different from homophily. The worst case bound (over all f) on the variance is in terms of the second eigenvalue, which is a measure of the uniform conductance over the entire network. A different nuanced interpretation of a network property from the variance becomes clearer when we consider the regular graph case. We can then rewrite the variance in terms of a measure of *homogeneity of neighbors*, as follows.

FACT 14. *For a function $f : V \to \{0,1\}$, and a d-regular graph G, define the* neighbor homogeneity *of f in G to be*

$$H(G,f) = \frac{1}{n} \sum_w |\{(u,v) \mid u,v \in N(w) \cap f\}|/d^2.$$

Then, the variance of Ideal *is given by* $\mathrm{var}(F) = H(G,f) - \bar{f}^2$.

For non-regular graphs, the corresponding homogeneity measure would weigh the nodes with the corresponding degrees.

5.5 Effect of rarity of truth

We then investigate the effect of the rarity (i.e., the quantity $\bar{f}$) on the error and the variance of the Ideal sampling strategy for the

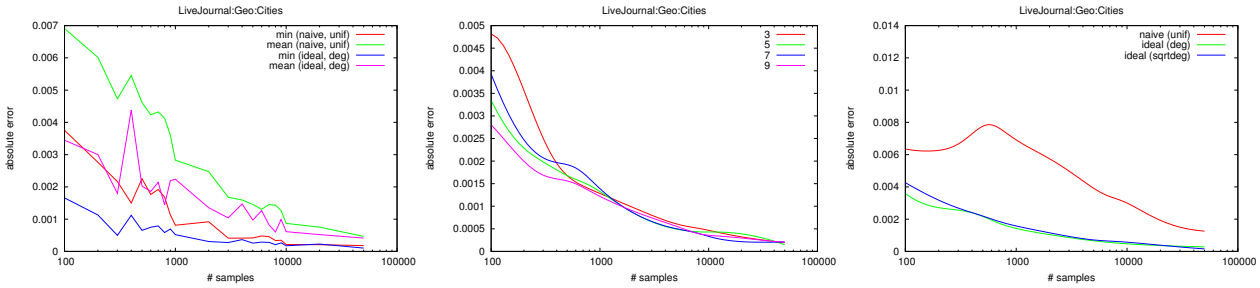

Figure 7: (a) Min vs mean in the partitioning of samples. (b) Effect of the sparsity parameter in Sparse on LIVEJOURNAL data. (c) Volume estimation.

uniform and degree bias. We bucket the cities based on the size of the attribute ($\bar{f}$). For each of these buckets, we computed the average error and the average variance for both the uniform and degree bias. The uniform has mostly a lower variance than the degree-biased samples, but the errors are comparable. The likely explanation is that the degree-biased sampling generates a few outliers, and the median operation helps reduce the error in spite of these outliers.

5.6 Minimum vs mean error of groupings

In order to construct the error values, we employed a median of means approach that we described earlier. Although this is commonly used in practice, there is no rule of thumb for deciding the right way to partition into groups in order to compute this median of means. We thus tried a number of combinations and reported the mean error over these partitions, which corresponds to the error if one chooses one of these partitions at random. Instead, we investigate how the error would look if we were to explicitly find out the partition with the minimum error and use that. Figure 7(a) shows the relation between the mean and the minimum computed over the partitions considered for each sample size, for the uniform and degree-biased Ideal strategy. Note that although the trends are similar for mean and minimum, there is a significant difference between the mean and minimum errors for each sample size. Our results have been reported using the mean aggregation: using the minimum prevents us from making statements such as "sampler X is better than sampler Y," since we would need to specify the partitioning strategy too. However, this leaves open the interesting question of how to define the optimal aggregation strategy when given a fixed number of samples.

5.7 Effect of the sparsity parameter

Figure 7(b) shows the effect of the sparsity parameter in the strategy Sparse, using the LIVEJOURNAL data. Interestingly sampling three neighbors is almost as good as sampling nine neighbors, indicating that a big savings is possible in terms of how many neighbors information we expect the respondents of the poll to summarize. A plausible explanation is that the expander properties of the network when sampling three neighbors on average are qualitatively equivalent to the properties arising from sampling nine neighbors.

5.8 Volume

Finally, we show (Figure 7(c) how our results also translate when we are trying to estimate volume of the attribute set, i.e., the fraction $d_f/2m$. For estimating volume, the Ideal strategies are orders of magnitude better than the Naive one.

6. CONCLUSIONS

We have analyzed a novel trade-off question: what are the benefits and difficulties that arise from sampling nodes of a social network and querying them about their neighbors' properties? We presented variance bounds of a variety of samplers that account for the network structure; our experiments strongly affirm the hypothesis that we started with, namely when done correctly (and incorporating a significant amount of knowledge about the network), the resulting estimators are considerably more efficient than standard polling. We believe that this work joins the exciting body of research at the intersection of statistics and social network theory.

7. REFERENCES

[1] L. Backstrom and J. M. Kleinberg. Network bucket testing. In *WWW*, pages 615–624, 2011.

[2] Z. Bar-Yossef, R. Kumar, and D. Sivakumar. Sampling algorithms: Lower bounds and applications. In *STOC*, pages 266–275, 2001.

[3] P. Boldi, F. Bonchi, C. Castillo, and S. Vigna. Voting in social networks. In *CIKM*, pages 777–786, 2009.

[4] R. Canetti, G. Even, and O. Goldreich. Lower bounds for sampling algorithms for estimating the average. *IPL*, 53(1):17–25, 1995.

[5] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Walking in Facebook: A case study of unbiased sampling of OSNs. In *INFOCOM*, pages 1–9, 2010.

[6] S. Goel and M. Salganik. Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in Medicine*, 28(17):2202–2229, 2009.

[7] O. Goldreich. A sample of samplers: A computational perspective on sampling. In O. Goldreich, editor, *Studies in Complexity and Cryptography*, volume 6650, pages 302–332. Springer, 2011.

[8] O. Goldreich and A. Wigderson. Tiny families of functions with random properties: A quality-size trade-off for hashing. *Random Struct. Algorithms*, 11(4):315–343, 1997.

[9] D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, pages 174–199, 1997.

[10] L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In *WWW*, pages 597–606, 2011.

[11] L. Katzir, E. Liberty, and O. Somekh. Framework and algorithms for network bucket testing. In *WWW*, pages 1029–1036, 2012.

[12] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD*, pages 631–636, 2006.

[13] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *IMC*, pages 390–403, 2010.

[14] D. Rothschild and J. Wolfers. Forecasting elections: Voter intentions versus expectations, 2011. Manuscript, `http://researchdmr.com/RothschildExpectations`.