Social Sampling

Marcel Mohler
mohlerm@student.ethz.ch

May 2, 2015

Abstract

This report provides the reader with an overview of the concept of *social sampling*. Social sampling is a class of methods where properties of a social network are used to obtain more accurate samplers than when using classical polling algorithms. The main interest lies in the correlation between the sampling size and the sampling error and the systematic bias due to the network structure.

I define the concept of a sampler and present a `Naive` as well as an `Ideal` sampling algorithm. Further I show and prove some of their most important properties regarding sample size and bias.

Eventually I present more realistic and efficient estimators and their real-world performance and demonstrate that social sampling is a powerful tool to obtain accurate estimates with very few samples.

1 Introduction

Polling is a widely used method to achieve a public opinion on a certain subject gained from a sample. For example, one could think of a poll executed in a social network asking whether or not someone has watched the soccer finals or polls before Bundesrat elections to estimate the outcome. Following the second example, pollsters would like to ask only a subset of the population while still getting significant results which party might win the vote. This is commonly known as sampling. The two fundamental quantities of interest in sampling are *number of samples* (how many citizens to ask) and the *error rate* (error caused by observing a sample instead of the whole voters). Chapter 2 will provide the reader with a detailed analysis of these two properties and how they relate in different sampling algorithms.

In practice we want to have a low error but still only require as few samples as possible. Since the sampling costs are usually a significant burden, research has proposed several alternatives to reduce the sample size. For instance instead of sampling uniformly one could choose its samples with a built in bias. However, as seen in chapter 2.3, this method is vulnerable to introduce a so called *systematic bias* which can lead to systematic errors in the outcome.

Another, recently popular approach is to use *expectation polling*, where voters are asked about their expected outcome of the poll. This stands in contrast to the classical *intent polling*, where voters are asked about their polling intent. David Rothschild and Justin Wolvers explored the value of expectation polling in a paper published 2009 [4]. Considering the focus of press and pollsters on intent polling they call it conventional wisdom that results obtained by intent polling are more accurate. Nonetheless, executed on the example of Presidential Electoral College races they provide robust evidence that expectation based polling yields to more accurate predictions of election outcomes. Unfortunately they fall short on providing any theoretical guarantees on either the sampling bias or sampling error.

What follows is a formal introduction to a group of estimators I call *samplers* as well as four concrete instances of these type of estimators, suggested and described by Anirban Dasgupta, Ravi Kumar and D Sivakumar in [2]. I start with a `Naive` sampler and demonstrate how to prove the relation between the error ϵ and sample size r in this simple example. Followed by an `Ideal` sampler where Dasgupta et al. incorporate the idea to reduce the sampling size by asking a member of a social network considering they will be able to summarize their friends opinions. The last two presented algorithms provide a more practical approach and try to model the method of expectation polling.

I analyze sampling bias and provide theoretical relations between sample size and sampling error including some proofs. I will also provide precise characterizations on errors and how they behave applied to large, real-world networks on the basis of Dasgupta's work.

2 Algorithms

2.1 Definitions

Note that the graph $G = (V, E)$ describes a social network with V denoting the set of nodes and E the set of edges. Further, $|V| = n$, $|E| = m$ and u, v are nodes of G. The set of neighbors of u is denoted by $N(u)$ and the degree of node u by $d_u = |N(u)|$. I will always assume $d_u \geq 1 \ \forall u$. Additionally, $f : V \to \{0, 1\}$ is a binary function which takes a node $u \in V$ as input and outputs whether this node satisfies a certain property or not and $\bar{f} = \frac{1}{n}|\{u \mid f(u) = 1\}|$ is the fraction I wish to estimate. The algorithms use S as the set of nodes called *sample* with $|S| = r$, the sample size. I will denote the probability of picking a node u with p_u for which holds $\sum_u p_u = 1$.

The pollsters goal is to achieve a good estimation of $\bar{f}$ using as few samples S as possible with respect to an additive error bound ϵ. I introduce the concept of a set of algorithms called sampler which is defined as follows:

Definition 2.1 (sampler). *A `sampler` is a randomized algorithm with input G (graph), r (sample size), ϵ (sampler error), δ (error probability), p (probability distribution) and oracle access to any function f which outputs an expectation $\hat{f}$ with probability $1-\delta$ and an error bounded by $|\hat{f}-\bar{f}| < \epsilon|$. Namely,*

$$\mathbb{P}[|\texttt{sampler}_f(G, r, \epsilon, \delta, p) - \bar{f}| > \epsilon] < \delta \ .$$

This is a similar definition of the one shown in Goldreichs work [3], extended to also include a graph and the probability distribution. δ is the probability that a sampler satisfies the wanted sampling error. In statistics $1-\delta$ is often called the *confidence* of the error. I do not put much emphasis on it in this work since I can set it arbitrarily low and considering I am mainly interested in comparing multiple samplers one would just choose the same δ for all sampler instances. Note that this definition of a sampler is in fact a special case of a statistical estimator. Therefore I can analyze a sampler for estimator properties like bias and I might also denote the samplers as estimators.

2.2 `Naive` sampler

I will start by looking at an intuitive approach for estimating $\bar{f}$ where one samples a set of nodes uniformly and polls each node u to test if $f(u) = 1$. The algorithm will return the fraction of nodes that satisfy the condition and I call this approach the `Naive` sampler presented in pseudo code in Alg. 1.

Alg. 1 Naive `sampler`$_f(G, r, \epsilon, \delta, p = 1/n)$

Input: Graph $G = (V, E)$, function $f : V \to \{0, 1\}$, sample size r
Output: $\hat{f} = \frac{1}{r}\sum_{u \in S} f(u)$

```
begin
    initialize $f^*$ with 0
    randomly draw a set $S$ with r samples from $V$ with
        probability $p_u = 1/n$ and with replacement
    for each $u \in S$ do
        $f^* = f^* + f(u)$
    end
return $\hat{f} = f^*/r$
end
```

Picking the samples uniformly means that each u is chosen with probability $p_u = 1/n$. Note that the drawing is done *with replacement* because considering $n >> r$ the possibility to draw a node twice is negligible and this has the effect of the drawings being independent from each other.

Theorem 2.1. *The* `Naive` *sampler with sampling probability* $p_u = 1/n$, *accuracy* ϵ *and confidence* $1 - \delta$ *requires* $\frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$ *sample nodes.*

Proof. Recall that I want to poll r uniformly chosen people independently and with replacement. The true fraction I want to approximate is $\bar{f}$. Let F_u be the random variable for $f(u) = 1$.

Since each of these drawings can be modeled by a coin flip with probabilities $\bar{f}$ and $1 - \bar{f}$ it follows that $F_u \sim \mathit{Bernoulli}(\bar{f})$ and $F_1, F_2, \ldots, F_r$ are independent. The expected value $\mathbf{E}[F_u] = \bar{f}$ therefore each pick unbiased. The `Naive` estimator is defined as $\hat{f} = \frac{1}{r} \sum_{u=1}^{r} F_u$. I want my estimate to have accuracy ϵ and confidence $1 - \delta$. This means that the probability that my estimation differs from the real value with a maximum of ϵ is larger than my defined confidence value (respectively $1 - \delta$):

$$\mathbb{P}[|\hat{f} - \bar{f}| \leq \epsilon] \quad \geq \quad 1 - \delta \ .$$

$\hat{f}$ multiplied by r is therefore a sum of Bernoulli trials and per definition $r\hat{f} \sim \mathit{Binomial}(r, \bar{f})$. For the expected value follows:

$$\mathbf{E}[r\hat{f}] \quad = \quad \mathbf{E}[r\frac{1}{r}\sum_{u=1}^{r} F_u] \quad = \quad \sum_{u=1}^{r} \mathbf{E}[F_u] \quad = \quad r\bar{f} \ .$$

I want to find a bound for the following expression:

$$\mathbb{P}[|r\hat{f} - \mathbf{E}[r\hat{f}]| \geq r\epsilon] \quad = \quad \mathbb{P}[|r\hat{f} - r\bar{f}| \geq r\epsilon] \quad = \quad \mathbb{P}[|\hat{f} - \bar{f}| \geq \epsilon] \ .$$

Using the two-sided Hoeffding's inequality which provides an upper bound for a sum of random variables derivating from their expected values the following holds for any $\epsilon \geq 0$

$$\mathbb{P}[|\hat{f} - \bar{f}| \geq \epsilon] \quad \leq \quad 2e^{-2r\epsilon^2} \ .$$

Remember I want the confidence of the estimator bounded by $1 - \delta$. I can also express this the other way around, which means the probability of an error larger than ϵ is less than δ. Therefore:

$$\begin{aligned} & \mathbb{P}[|\hat{f} - \bar{f}| > \epsilon] \quad \leq \quad \delta \\ \Leftrightarrow \quad & 2e^{-2r\epsilon^2} \quad \leq \quad \delta \\ \Leftrightarrow \quad & e^{2r\epsilon^2} \quad \geq \quad \frac{2}{\delta} \\ \Leftrightarrow \quad & 2r\epsilon^2 \quad \geq \quad \ln \frac{2}{\delta} \\ \Leftrightarrow \quad & r \quad \geq \quad \frac{1}{2\epsilon^2} \ln \frac{2}{\delta} \ . \end{aligned}$$

This matches the bound of Theorem 2.1 and concludes the proof. □

Dasgupta et al. [2] proposed $r = 2/(\epsilon^2\delta)$ and since $r = 2/(\epsilon^2\delta) \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$ this Lemma follows directly:

Lemma 2.2. *Using $r = 2/(\epsilon^2\delta)$ samples, with probability $1 - \delta$ the `Naive` sampler will give an estimate $\hat{f}$ such that $|\hat{f} - \bar{f}| < \epsilon$.*

2.3 `Ideal` sampler

Now, I want to improve the previous estimator by using the concept of social sampling. So basically when I poll a node u I expect to get an estimation of $f(v) \mid v \in N(u)$ to decrease the number of samples needed. I call this sampler `Ideal`[2] and the pseudo-code is shown in Alg. 2.

Alg. 2 Ideal `sampler`$_f(G, r, \epsilon, \delta, p)$

Input: Graph $G = (V, E)$, function $f : V \to \{0, 1\}$, sample size r, distribution p
Output: $\hat{f} = \frac{1}{nr} \sum_{u \in S} \frac{1}{p_u} \sum_{v \in N(u)} f(v)/d_v$

```
begin
    initialize $f^*$ with 0
    randomly draw a set $S$ with r samples from $V$ with
        probability $p_u$ and with replacement
    for each $u \in S$ do
        for each $v \in N(u)$ do
            $f^* = f^* + \frac{1}{np_u} f(v)/d_v$
        end
    end
    return $\hat{f} = \frac{f^*}{r}$
end
```

Before looking at sample size and error I want to explain the scaling of each $f(v)$ with $1/d_v$ on a small example. As briefly mentioned in the introduction the network structure is of great importance to the sampling method. Consider a graph structure as shown in Figure 1 where a single `RED` node is connected to several `BLUE` nodes. Imagine a sampler that samples the blue nodes and for each includes the value of their neighbors. Even though the portion of `RED` nodes is way smaller, the result is very biased towards the decision of RED since it is included in the neighbor set of each sampled node. So a sampler dividing each of the neighbors values by it's degree undoes that bias. The factors $1/n$ and $1/p_u$ are needed because I now want to consider different, non-uniform p_us compared to only uniform ones in the `Naive` sampler. This allows me to for for example to prefer nodes with

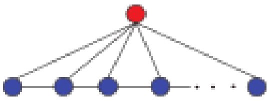

Figure 1: neighbor scaling

more neighbors by giving them a higher sampling probability. The scaling itself will be more clear when looking at the expected value in Lemma 2.3 and the bias property of the sampler.

I am going to provide a proof that this sampler is indeed unbiased and again present bounds on the sampling size. First and foremost, I introduce A as adjacency matrix of the graph G, which means $A_{uv} = 1$ if and only if $(u, v) \in E$. D is called the diagonal matrix such that $D_{uu} = d_u$. Let the diagonal matrix P be $P_{uu} = p_u$ with $\sum_u p_u = 1$ for any vector $p \in \mathbb{R}^n$. I also introduce $\mathbb{F}$ which is the indicator (column) vector of the set $u \mid f(u) = 1$, i.e., $\mathbb{F}_u = f(u)$. $\mathbb{1}$ is simply an n-element vector of all 1's. $F_u = \frac{e_u^T AD^{-1}\mathbb{F}}{np_u}$ is a random variable such that $F_u = f(u)$. To help understand this expression, note that vth entry of $D^{-1}f$ are the neighbors choices scaled with their degree and the uth entry of $AD^{-1}f$ is the respective sum $\sum_{v\in N(u)} f(v)/d_v$. The `Ideal` sampler is defined as $\hat{f} = \frac{1}{r}\sum_u F_u$.

Lemma 2.3. *The random variable F_u satisfies $E[F_u] = \bar{f}$ and $E[F_u^2] = \frac{1}{n^2}\mathbb{F}^T D^{-1}AP^{-1}AD^{-1}\mathbb{F}$.*

Proof.

$$\begin{aligned}
\mathbf{E}[F_u] &= \sum_u p_u \frac{e_u^T AD^{-1}\mathbb{F}}{np_u} \quad = \quad \sum_u e_u^T AD^{-1}\mathbb{F}/n \\
&= \mathbb{1}^T AD^{-1}\mathbb{F}/n \quad = \quad \mathbb{1}^T\mathbb{F}/n \quad = \quad \bar{f} \\
\mathbf{E}[F_u^2] &= \sum_u p_u \left(\frac{e_u^T AD^{-1}\mathbb{F}}{np_u}\right)^2 \quad = \quad \frac{1}{n^2}\sum_u \frac{(e_u^T AD^{-1}\mathbb{F})^2}{p_u} \\
&= \frac{1}{n^2}\sum_u \mathbb{F}^T D^{-1}Ae_u e_u^T AD^{-1}\mathbb{F}/p_u \\
&= \mathbb{F}^T D^{-1}A\Big(\sum_u e_u e_u^T/p_u\Big)AD^{-1}\mathbb{F}/n^2 \\
&= \mathbb{F}^T D^{-1}AP^{-1}AD^{-1}\mathbb{F}/n^2
\end{aligned}$$

□

Since the expected value of F_u is $\bar{f}$ the estimator is indeed unbiased. This would not hold if each $f(v)$ is not scaled by $1/np_u$.

I still want to find a bound for the number of samples, similar to the `Naive` sampler. Therefore I first provide a bound for the variance in the special case where nodes u are sampled with probability proportional $|N(u)| =$

d_u, meaning $p_u = \frac{d_u}{2m}$. As mentioned exemplary before this prefers high degree nodes and allows us to get an estimation of more nodes by using fewer samples. The division by $2m$ is needed to achieve $\sum_u p_u = 1$ because the sum of all degrees equals twice the total amount of edges.

Lemma 2.4. *Let the sampling probability be $p_u = \frac{d_u}{2m}$. Then variance $\textbf{Var}(F_u) \leq (2m/n^2)\lambda_2^2 ||D^{-1/2}f||^2$ with λ_2 being the second largest eigenvalue of matrix L defined as $L = D^{-1/2}AD^{-1/2}$.*

Proof.

$$\begin{aligned}
\textbf{Var}(F_u) &= \textbf{E}[F_u^2] - (\textbf{E}[F_u])^2 \quad = \quad \mathbb{F}^T D^{-1} A P^{-1} A D^{-1} \mathbb{F}/n^2 - (\mathbb{F}/n)^2 \\
&= \mathbb{F}^T D^{-1} A (2mD^{-1}) A D^{-1} \mathbb{F}/n^2 - (\mathbb{F}/n)^2 \\
&= \frac{2m}{n^2} \mathbb{F}^T D^{-1} A D^{-1} A D^{-1} \mathbb{F} - (\mathbb{F}/n)^2 \\
&= \frac{2m}{n^2} \mathbb{F}^T D^{-1/2} L^2 D^{-1/2} \mathbb{F} - \mathbb{F}^T \mathbb{1}\mathbb{1}^T \mathbb{F}/n^2 \\
&= \frac{2m}{n^2} \mathbb{F}^T D^{-1/2} \left(L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m} \right) D^{-1/2} \mathbb{F} \\
&\leq \frac{2m}{n^2} \left|\left| L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m} \right|\right| ||D^{-1/2}\mathbb{F}||^2
\end{aligned}$$

In fact the recently defined matrix L has interesting properties. For a graph G one can compute the Laplacian matrix $\mathcal{L}$ as $\mathcal{L} = D - A$ where A is the adjacency matrix and D the degree matrix as defined above. Taking $I - L$ is called the normalized Laplacian. Because of this association it can be shown that the second largest eigenvalue of L^2 equals $\lambda_2^2 = ||L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m}||$ and therefore $1 - \lambda_2$ is the second smallest eigenvalue of the normalized Laplacian. The proof follows. □

I can now use this bound and the Chebyshev inequality to prove the following bound on r.

Theorem 2.5. *Using $r = \frac{2\textbf{Var}(F_u)}{\epsilon^2\delta}$ and $p_u = \frac{d_u}{2m}$, with probability $1 - \delta$ the `Ideal` sampler will give an estimate $\hat{f}$ such that $|\hat{f} - \bar{f}| < \epsilon$.*

Proof.

$$\begin{aligned}
\mathbb{P}[|\hat{f} - \bar{f}| > \epsilon] \quad &\leq \quad \frac{2\textbf{Var}(\hat{f})}{\epsilon^2} \quad \leq \quad \frac{2\textbf{Var}(\hat{F})}{r\epsilon^2} \\
&\leq \quad \frac{2\textbf{Var}(F_u)\epsilon^2\delta}{2\textbf{Var}(F_u)\epsilon^2} \quad \leq \quad \delta
\end{aligned}$$

□

There are further, lower bounds of the sample size if G is d-regular and especially if the graph is an expander. Anyway, this requires a detailed look at the graphs properties and further investigations and is considered out of the scope of this report [1].

2.4 `Sparse` sampler

The `Ideal` estimator samples *all* neighbors of the sampled node. Thinking back to social network this means a person is expected to represent all of its neighbors choices. Now, I present a variant of this estimator that only samples a subset T_u of u's neighbors. This models a scenario where either polling all neighbors is expensive respectively impossible (e.g. API limitations) or where we use further heuristics to select an arbitrary amount of neighbors. In the second case one could think of a poll like "*Think of three of your female friends and tell us how many are married.*".

This sampler uses an additional parameter called k, which is the size of set T'_u, a randomly chosen subset from T_u. I am aware this contradicts the initial definition of a sampler but since k has no usage in the other samplers I keep the original definition anyways while still calling the `Sparse` estimator a sampler.

In this implementation I will pick each neighbor with probability inversly proportional to their degrees and pick the subset of k neighbors with probability $1/|T_u|$. A pseudo code implementation is presented in Alg. 3.

Providing a bound on variance and sample size is considered harder and not shown in this report.

2.5 `Expectation` sampler

The final presented sampler, short `Expec`, is based on the research and ideas of Rothschild and Wolfers in [4]. They showed with statistical measurements that using a method called *expectation polling* can lead to significantly better estimations than the classical *intent polling*. In practise this means posing a "*Which party will win the next vote?*" instead of "*Who do you vote for?*". They base this on the assumption that a respondent often gives an aggregate view of its friends and family (i.e. neighbors in a social network). The presented algorithm Alg. 4, which aims to implement expectation polling lets each sampled node u report a fraction $G(u)$ instead of a simple 0 or 1. $G(u)$ represents the estimated fraction of the neighbors with $f(v) = 1$ and has values r_u (see Alg. 4) or 0.

Compared with the `Ideal` sampler there exist a similar bound on r and we present this without a proof:

Theorem 2.6. *Using* $r = \frac{4m\lambda_2^2||D^{-1/2}\mathbb{F}||||D^{-1/2}(\mathbb{1}-\mathbb{F})||}{n^2\epsilon^2\delta}$, *with probability* $1 - \delta$ *the `Expec` sampler will give an estimate* $\hat{f}$ *such that* $|\hat{f} - \bar{f}| < \epsilon$.

Alg. 3 Sparse $\texttt{sampler}_f(G, r, \epsilon, \delta, p, k)$

Input: Graph $G = (V, E)$, function $f : V \rightarrow \{0, 1\}$, sample size r, distribution p, parameter k

Output: $\hat{f} = \frac{1}{nr} \sum_{u \in S} \frac{1}{p_u} (|T_u|/k) \sum_{v \in T'_u} f(v)$

```
begin
    initialize $f^*$ with 0
    randomly draw a set $S$ with r samples from $V$ with
        probability $p_u$ and with replacement
    for each $u \in S$ do
        1. randomly draw a set $T_u$ by picking each neighbor
            $v \in N(u)$ with probability $1/d_v$
        2. randomly draw a set $T'_u \subseteq T_u$ by picking
            $k$ elements of $T_u$ without replacement with
            probability $1/|T_u|$.
        for each $v \in T'_u$ do
            $f^* = f^* + \frac{1}{np_u}(|T_u|/k) f(v)$
        end
    end
    return $\hat{f} = \frac{f^*}{r}$
end
```

3 Results

3.1 Setup

Dasgupta et al. [2] performed various experiments to compare the performance of the presented samplers. They focused on two large, publicly available databases `LiveJournal` and `DBLP`. Note that the graphs in the following chapter are directly copied from their published paper.

The `LiveJournal` network contains 5.36M nodes and approximately 160M edges and is based on a snapshot of the social network `http://livejournal.com` in March 2008. In addition to the network they extracted the users' age, location (city, state, country) and a list of interests where available on the users' public profiles. They used additional filter techniques to remove invalid entries, lowering the amount of nodes to 40%-60% per attribute.

The `DBLP` network contains 368K nodes and 8M edges and consists of the content of the DBLP database located at `http://www.informatik.uni-trier.de/~ley/db/`. The nodes stand for authors while the edges represent co-authorship on some publication. The node attributes are formed

Alg. 4 Expec $\texttt{sampler}_f(G, r, \epsilon, \delta, p)$

Input: Graph $G = (V, E)$, function $f : V \to \{0, 1\}$, sample size r, distribution p

Output: $\hat{f} = \frac{1}{nr} \sum_{u \in S} \frac{1}{p_u} G(u)$

```
begin
    initialize $f^*$ with 0
    randomly draw a set $S$ with r samples from $V$ with
        probability $p_u$ and with replacement
    for each $u \in S$ do
        1. compute $q_u = \sum_{v \in N_u} \frac{f_v}{d_v}$
        2. compute $r_u = \sum_{v \in N_u} \frac{1}{d_v}$
        3. flip coin with probability $\frac{q_u}{r_u}$
        if heads
            $f^* = f^* + r_u$
        else
            $f^* = f^* + 0$
        end
    end
return $\hat{f} = \frac{f^*}{nr}$
end
```

by the authors publication venues, however only a random set of 100 venues with probability proportional to their popularity were considered.

Since the samplers have been designed keeping the Definition 2.1 in mind, recall that they provide an additive ϵ-approximation. So most importantly Dasgupta et al. measure the absolute error of each sampler with respect to the true value. If an estimator outputs $\hat{f}$ approximating $\bar{f}$ they measure the error as $|\bar{f} - \hat{f}|$.

To get a relation between error and sample size they vary the sample size r from a minimum of $r = 100$ to $r = 50000$. They use the distributions mentioned in chapter 2: uniform at random (`unif`), proportional to the degree of the node (`deg`) and additionally, proportional to the square root of the degree of the node (`sqrtdeg`). Note that in case of the `Sparse` sampler $k = 5$ has been used. They omit any other combinations of algorithm and distributions since they almost always did not result in better performance and are computationally more expensive. The results were achieved using a median of means approach. This is commonly used when doing these kinds of measurements and works in the following way: one creates l partitions called $A_1, ..., A_l$ of the set of samples A. l is a parameter and in this case l =1,3,5,7 and 9 have been used. Next, for each l the median is calculated from applying the estimator to A_1 until A_l. The final result and the one presented in the graphs is the mean of these medians from all l. This considers choosing partitions at random and prevents the need to specify the partition strategy when comparing the different samplers.

3.2 Performance on LiveJournal

For analyzing the performance on LiveJournal they are interested in a set of five cities, five states, five countries, a set of ages (20, 25, 30, 35 and 40), age-buckets ([20,29),...,) and a set of interests (various sports) each with a typical size ranging from 0.4% to 4%. The correlation between number of samples and absolute error for the presented samplers is displayed in Figure 2.

As expected, in all scenarios, the error decreases faster than linearly with an increased number of samples. Generally `deg` and `sqrtdeg` show similar performance and I will not discuss the details in this paper. To get a feeling on r in relation to the whole network, keep in mind that when sampling 10000 nodes one is already sampling 2% (on average) of the whole network's nodes, an already significant sample size.

In all attributes the `Ideal` sampler shows the best performance, for example in cities until about 10000 samples its error is about half as `Naive`'s. `Naive` generally performs better than `Expec` but surprisingly at states and countries it performs similar to the `Ideal` estimator. The `Sparse` sampler is usually similar to the `Naive` approach with being better at cities and worse at the states attribute. Most interestingly `Expec` shows a very indifferent performance. Out of all algorithms it seems to have the highest non-monotone behavior at around $r = 200$. It is assumed that this is due to `Expec` having the highest variance because of the discretization of the value returned by each sampled node.

3.3 Performance on DBLP

In DBLP they are interested in 100 venues of publication ranging in size from 0.01% to 3% in size. The correlation between sample size and absolute error is drawn in Figure 3.

The improvements offered by `Ideal` over `Naive` are even larger. In comparison the `Naive` sampler requires at least 3-5 times more samples than *Ideal* for achieving the same error.

3.4 Effect of sparsity parameter

When designing the `Sparse` sampler Alg. 3 I introduced a new parameter k which I call the sparsity parameter since it defines the amount of the picked neighbors. The measurements displayed in Figure 4 show that using $k = 9$

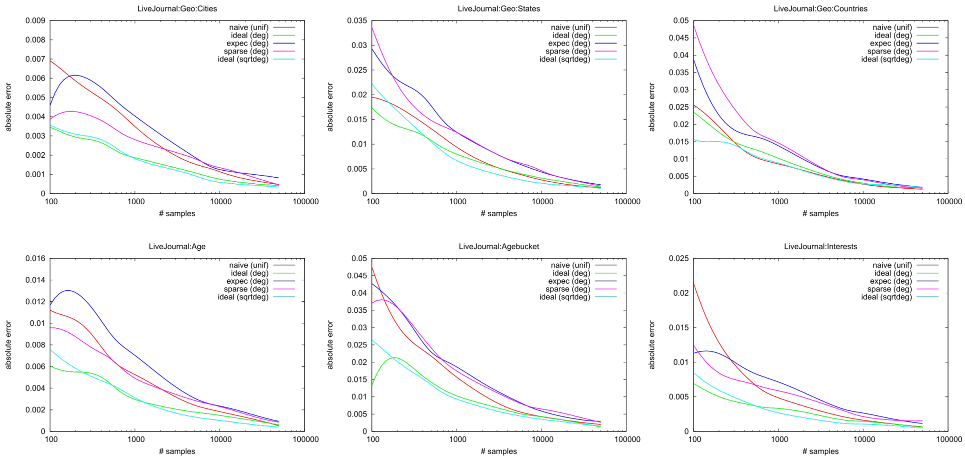

Figure 2: #samples vs error of various samplers on the LiveJournal data [2]

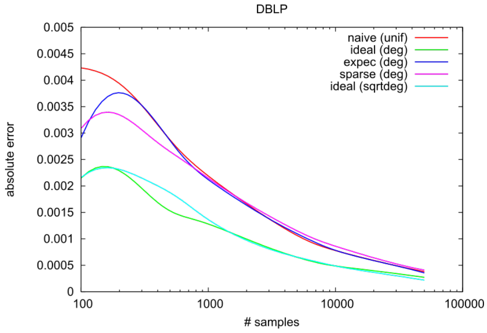

Figure 3: #samples vs error of various samplers on the DBLP data [2]

does not lower the error significantly, especially considering sample sizes above 400. The authors of these experience and me assume this is related to the fact that the network has strong connectivity properties.

4 Conclusion

In this report I introduced the reader to the concept of a sampler and analyzed if polling a social network can result in less samples being needed. I presented various different sampling algorithms including bounds on their sample size to guarantee a certain error. The measurements showed that when executed correctly, meaning choosing an appropriate algorithm, consider a significant knowledge of the underlying network graph, results in considerably less samples needed compared to a naive sampling approach.

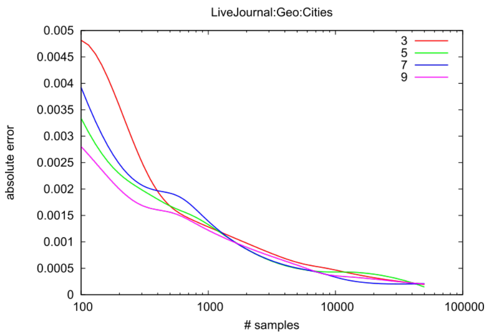

Figure 4: #samples vs error with different sparsity parameters [2]

References

[1] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986.

[2] A. Dasgupta, R. Kumar, and D. Sivakumar. Social sampling. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 235–243, 2012.

[3] O. Goldreich. A sample of samplers - A computational perspective on sampling (survey). *Electronic Colloquium on Computational Complexity (ECCC)*, 4(20), 1997.

[4] D. Rothschild. Forecasting elections comparing prediction markets, polls, and their biases. *Public Opinion Quarterly*, 73(5):895–916, 2009.