Social Sampling

Marcel Mohler
mohlerm@student.ethz.ch

April 8, 2015

Abstract

This report provides the reader with an overview of the concept of *social sampling*. Social sampling is a class of methods where informants in a poll answer with a summary of their friends cumulative responses. The main interest lies in the correlation between the sampling size and the sampling error and the systematic bias due to the network structure.

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

Eventually we 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 polls before *Bundesrat* elections to estimate the outcome or polls to receive student's opinions on their visited lectures. Following the first example, pollers would like to ask only a subset of the population while still get significant results which party might win the vote. This is commonly known as sampling. The two fundamental quantities of interest in polling are *number of samples* (how many citizens to ask) and the *error rate* (error caused by observing a sample instead of the whole voters). They are related in a way, that if we wish to approximate a fraction within an additive error of ϵ, roughly $O(1/\epsilon^2)$ samples are both necessary and sufficient **[[TODO: cite]]**.

In practice we want to have a low error but still only require as little 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 their samples with a built in bias. However this method is vulnerable to introduce a so called *systematic bias* which can lead to systematic errors in the outcome.

1

Another, recently popular approach is to use "expectation polling", where voters are asked about their expectation about the outcome of the poll in contrast to the more classical "intent polling" (where voters are asked about their polling intent). In [3] David Rothschild and Justin Wolvers explored the value of expecation polling. Considering the focus of press and pollsters on intent polling they call it conventional wisdom that these type of polls are more accurate. However, executed on the example of Presidential Electoral College races they provide robust evidence that expectation based polling yields to more accurate prediction of election outcomes. Unfortunately they fall short on providing any theoretical guarantees on either the sampling bias or sampling error.

In this paper we present an extended idea of "expectation polling" called "social sampling", suggested and described by Anirban Dasgupta, Ravi Kumar and D Sivakumar in [1]. Their idea is to reduce the sampling size by asking a member of a social network considering they will be able to summarize their friends opinions when asked the right questions. Even though this means that the actual structure of the network plays a role when analysing the error, they assume a saving in number of samples by a factor of around d, the average amount of friends of a member in a social network.

What follows are various social sampling strategies with the aim to observe sampling bias and sampling error. We will also provide precise characterizations on errors and how they behave applied to large, real-world graphs.

2 Algorithms

2.1 Definitions

Note that 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 $d_u = |N(u)|$. We will always assume $d_u \geq 1 \forall u$. Additionally, $f : V \to \{0, 1\}$ is a binary function which takes a node 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 we wish to estimate. The algorithms use S as the set of nodes called "sample" with $|S| = r$ (sample size). We will denote the probability distribution by a a vector $p \in \mathbb{R}^n$ with $\sum_u p_u = 1$.

Our goal is to estimate the portion of nodes $v \in C$ such that $f(v) = 1$. We introduce the concept of a set of algorithms called sampler which is defined as follows:

Definition 2.1 (sampler) *A sampler $\hat{f}(n, \epsilon, \delta)$ is a randomized algorithm with input r (sample size), ϵ (sampler accuracy), δ (sampler error), p (prob-*

ability distribution) and function f which outputs an expectation $\hat{f}$ with probability $1 - \delta$ and $|\hat{f} - \bar{f} < \epsilon|$. Namely,

$$\boldsymbol{Pr}[|\hat{f}(n, \epsilon, \delta, p) - \bar{f}| > \epsilon] < \delta$$

Note, that this definition of a sampler is in fact a special case of a statistical estimator. Therefore we can analyze our sampler for estimator properties like bias and we will use both expressions equivalently.

2.2 Naive sampler

We will start by looking at an intuitive approach for estimating $\bar{f}$ where we sample a set of nodes and poll each node u to test if $f(u) = 1$. The algorithm will return the fraction of nodes that satisfy the condition and we call this approach the `Naive` estimator [2].

Alg. 1 Naive size sampler(G, f, r, p)

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

1	**begin**
2	initialize f^* with 0
3	randomly draw a set S with r samples from V with probability p_u and with replacement
4	**for** each $u \in S$ **do**
5	$f^* = f^* + f(u)$
6	**end**
7	**return** $\hat{f} = f^*/r$
8	**end**

In the analysis we use a uniform distribution p where we pick each node u with probability $p_u = 1/n$.

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

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

It follows that $F_u \sim Bernoulli(\bar{f})$ and $F_1, F_2, \ldots, F_r$ are independent. Our `Naive` estimator is defined with $F = \sum_{u=1}^{r} F_u$ as $\hat{f} = F/r$. We want our estimate to have accuracy ϵ and confidence $1 - \delta$:

$$\mathbf{Pr}[|\hat{f} - \bar{f}| \leq \epsilon] \quad \geq \quad 1 - \delta$$

Since $F \sim Binomial(r, \bar{f})$, it follows

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

Using the two-sided Chernoff Bounds the following inequality holds for any $\gamma \geq 0$

$$\mathbf{Pr}[|F - r\bar{f}| \geq \gamma r\bar{f}] \quad \leq \quad 2\exp(-\frac{\gamma^2}{2+\gamma} \cdot r\bar{f})$$

$$\Leftrightarrow \quad \mathbf{Pr}[|\hat{f} - \bar{f}| \geq \bar{f}\gamma] \quad \leq \quad 2\exp(-\frac{\gamma^2}{2+\gamma} \cdot r\bar{f})$$

To achieve $\hat{f}$ bounded by ϵ we set $\epsilon = \gamma \bar{f}$, so $\gamma = \epsilon/\bar{f}$. Inserted in the formula above:

$$\mathbf{Pr}[|\hat{f} - \bar{f}| \geq \epsilon] \quad \leq \quad 2\exp(-\frac{\epsilon^2/\bar{f}^2}{2+\epsilon/\bar{f}} \cdot r\bar{f}) \quad = \quad 2\exp(-\frac{\epsilon^2}{2\bar{f}+\epsilon} \cdot r)$$

Since the largest possible value of $\bar{f}$ is 1, because all nodes u have $f(u) = 1$, the term in $\exp(\cdot)$ is an upper bound

$$\frac{\epsilon^2}{2\bar{f}+\epsilon} \quad \geq \quad \frac{\epsilon^2}{2+\epsilon}$$

and therefore

$$\mathbf{Pr}[|\hat{f} - \bar{f}| \geq \epsilon] \quad \leq \quad 2\exp(-\frac{\epsilon^2}{2+\epsilon} \cdot r)$$

Remember we want the confidence of the estimator bounded by $1 - \delta$, this means

$$2\exp(-\frac{\epsilon^2}{2+\epsilon} \cdot r) \leq \delta$$

$$\Leftrightarrow \quad \exp(\frac{\epsilon^2}{2+\epsilon} \cdot r) \quad \geq \quad \frac{2}{\delta}$$

$$\Leftrightarrow \quad \frac{\epsilon^2}{2+\epsilon} \cdot r \quad \geq \quad \ln\frac{2}{\delta}$$

$$\Leftrightarrow \quad r \quad \geq \quad \frac{2+\epsilon}{\epsilon^2}\ln\frac{2}{\delta}$$

Since $\frac{2+\epsilon}{\epsilon^2}\ln\frac{2}{\delta}$ lies within $O(\frac{1}{\epsilon^2})$ this concludes the proof. $\qquad\square$

Dasgupta et al. [1] proposed $r = 2/(\epsilon^2\delta)$ and since $r = 2/(\epsilon^2\delta) \geq \frac{2+\epsilon}{\epsilon^2}\ln\frac{2}{\delta}$ 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`

We now want to improve our estimator by using the concept of social polling, so basically when we poll a node u we expect to get an estimation of $f(v) \mid v \in N(u)$ to decrease the number of samples needed. As briefly mentioned in the introduction the network structure is of great importance to the sampling method. We will illustrate this on a small example. Consider a graph structure as shown in Figure 1 where a single `RED` node is connected to several `BLUE` nodes. Even though the portion of `RED` nodes is way smaller, using the `Naive` estimator the result is very biased towards the decision of RED.
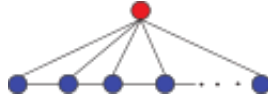

Figure 1: neighbor scaling

We introduce a social polling sampler called `Ideal` sampler that circumvents this bias by dividing each $f(v)$ by d_v. The pseudocode is shown in Alg. 2. We are going to provide a proof that this sampler is indeed unbiased and again present bounds on the sampling size.

Alg. 2 Ideal size sampler(G, f, r, p)

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

Output: $\hat{f} = \frac{1}{nr} \sum_{u \in S} \frac{1}{p_u} \sum_{v \in N(u)} f(v)/d_v$

   1   **begin**
   2   initialize f^* with 0
   3   randomly draw a set S with r samples from V with probability p_u and with replacement
   4   **for** each $u \in S$ **do**
   5   **for** each $v \in N(u)$ **do**
   6               $f^* = f^* + \frac{1}{p_u} f(v)/d_v$
   7   **end**
   8   **end**
   9   **return** $\hat{f} = \frac{f^*}{nr}$
  10   **end**

First and foremost, we introduce A as adjacency matrix of the graph G, which means $A_u v = 1$ if and only if $(u, v) \in E$. D is called the diagonal ma-

trix such that $D_{uu} = d_u$. Let the diagonal matrix P be $P_{uu} = p_u$ with $\sum_u p_u$ for any vector $p \in \mathbb{R}^n$. We 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. Our `Ideal` sampler is defined with $F = \sum_u \frac{e_u^T A D^{-1} \mathbb{F}}{n p_u}$ as $\hat{f} = F/r$.

Lemma 2.3 *The random variable F satisfies $E[F] = \mathbb{F}/n$ and $E[F^2] = \frac{1}{n^2} \mathbb{F}^T D^{-1} A P^{-1} A D^{-1} \mathbb{F}$.*

Proof:

$$
\begin{aligned}
\mathbf{E}[F] &= \sum_u p_u \frac{e_u^T A D^{-1} \mathbb{F}}{n p_u} \quad = \quad \sum_u e_u^T A D^{-1} \mathbb{F}/n \\
&= \mathbb{1}^T A D^{-1} \mathbb{F}/n \quad = \quad D D^{-1} \mathbb{F}/n \quad = \quad \mathbb{F}/n \\
\mathbf{E}[F^2] &= \sum_u p_u \left(\frac{e_u^T A D^{-1} \mathbb{F}}{n p_u} \right)^2 \quad = \quad \frac{1}{n^2} \sum_u \frac{(e_u^T A D^{-1} \mathbb{F})^2}{p_u} \\
&= \frac{1}{n^2} \sum_u \mathbb{F}^T D^{-1} A e_u e_u^T A D^{-1} \mathbb{F}/p_u \\
&= \mathbb{F}^T D^{-1} A \left(\sum_u e_u e_u^T/p_u \right) A D^{-1} \mathbb{F}/n^2 \\
&= \mathbb{F}^T D^{-1} A P^{-1} A D^{-1} \mathbb{F}/n^2
\end{aligned}
$$

Since $\mathbb{F}/n$ is a vectorial way of expressing the fraction $\bar{f}$ the estimator is indeed unbiased.

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

Lemma 2.4 *Let the sampling probability be $p_u = \frac{d_u}{2m}$. Then $\mathbf{var}(F) \leq (2m/n^2)\lambda_2^2 \|D^{-1/2} f\|^2$ with λ_2 being the second largest eigenvalue of matrix $L = D^{-1/2} A D^{-1/2}$.*

Proof:

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

One can show by using linear algebra that $\lambda_2^2 = ||L^2 - \frac{D^{1/2}\mathbb{1}\mathbb{1}^T D^{1/2}}{2m}||$.

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

Theorem 2.5 *Using $r = \frac{2var(F)}{\epsilon^2 \delta}$ 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:

$$
\begin{aligned}
\mathbf{Pr}[|\hat{f} - \bar{f}| > \epsilon] &\leq \frac{2\mathbf{var}(\hat{f})}{\epsilon^2} &\leq \frac{2\mathbf{var}(\hat{F})}{r\epsilon^2} \\
&\leq \frac{2\mathbf{var}(F)\epsilon^2\delta}{2\mathbf{var}(F)\epsilon^2} &\leq \delta
\end{aligned}
$$

2.4 Sparse sampler

The `Ideal` sampler samples all neighbors of the sampled node. Thinking back to social network this means a person is expected to represent all of their neighbors choices. Now, we present a variant of this estimator that only samples a subset T_u of u's neighbors. This represents 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.*".

Be aware that this sampler uses an additional parameter called k. This is the size of set T'_u, a randomly chosen subset from T_u. We are aware this contradicts the initial definition of a sampler but we since k has no usage in the other samplers we will stick to this defintition.

In this implementation we will pick each neighbor with probability inversly proportional to their degrees and pick the subset of k neighbors with probability $1/|T_u|$.

There are no proofs on the bound of the sample size yet and deriving one is considered out of the scope of this paper.

Alg. 3 Sparse size sampler(G, f, r, p, k)

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

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

1	**begin**
2	initialize f^* with 0
3	randomly draw a set S with r samples from V with probability p_u and with replacement
4	**for** each $u \in S$ **do**
5	1. randomly draw a set T_u by picking each neighbor $v \in N(u)$ with probability $1/d_v$
6	2. randomly draw a set $T'_u \subseteq T_u$ by picking k elements of T_u without replacement with probability $1/|T_u|$.
7	**for** each $v \in T'_u$ **do**
8	$f^* = f^* + \frac{1}{p_u}(|T_u|/k)f(v)$
9	**end**
10	**end**
11	**return** $\hat{f} = \frac{f^*}{nr}$
12	**end**

2.5 Expectation sampler

The final presented sampler is based on the research of Rothschild and Wolfers in [3]. 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 that aims to implement expectation polling let each sampled node u report a fraction instead of a simple 0/1. The fraction represents the estimatated fraction of the neighbors with $f(v) = 1$. 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 `Expectation` sampler will give an estimate* $\hat{f}$ *such that* $|\hat{f} - \bar{f}| < \epsilon$.

Alg. 4 Expec winner sampler(G, f, r, p)

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

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

1	**begin**
2	initialize f^* with 0
3	randomly draw a set S with r samples from V with probability p_u and with replacement
4	**for** each $u \in S$ **do**
5	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 $\frac{q_u}{r_u}$
6	**if** heads
7	$f^* = f^* + r_u$
8	**else**
9	$f^* = f^* + 0$
10	**end**
11	**end**
12	**return** $\hat{f} = \frac{f^*}{nr}$
13	**end**

3 Results

References

[1] A. Dasgupta, R. Kumar, and D. Sivakumar. Social sampling. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 235–243. ACM, 2012.

[2] O. Goldreich. A sample of samplers: A computational perspective on sampling. *def*, 1:2n, 1997.

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