Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Partitioned Sampling of Public Opinions Based on Their Social Dynamics*

Weiran Huang
IIIS, Tsinghua University
Beijing, China
huang.inbox@outlook.com

Liang Li
AI Department, Ant Financial Group
Hangzhou, Zhejiang, China
liangli.ll@alipay.com

Wei Chen
Microsoft Research
Beijing, China
weic@microsoft.com

Abstract

Public opinion polling is usually done by random sampling from the entire population, treating individual opinions as independent. In the real world, individuals' opinions are often correlated, e.g., among friends in a social network. In this paper, we explore the idea of partitioned sampling, which partitions individuals with high opinion similarities into groups and then samples every group separately to obtain an accurate estimate of the population opinion. We rigorously formulate the above idea as an optimization problem. We then show that the simple partitions which contain only one sample in each group are always better, and reduce finding the optimal simple partition to a well-studied Min-r-Partition problem. We adapt an approximation algorithm and a heuristic to solve the optimization problem. Moreover, to obtain opinion similarity efficiently, we adapt a well-known opinion evolution model to characterize social interactions, and provide an exact computation of opinion similarities based on the model. We use both synthetic and real-world datasets to demonstrate that the partitioned sampling method results in significant improvement in sampling quality and it is robust when some opinion similarities are inaccurate or even missing.

1 Introduction

Public opinion is essential nowadays for governments, organizations and companies to make decisions on their policies, strategies, products, etc. The most common way to collect public opinions is polling, typically done by randomly sampling a large number of individuals from the entire population and then interviewing them by telephone. This naive method is unbiased, but conducting interviews is very costly. On the other hand, in recent years, more and more online social media data are available and have been used to predict public opinions on certain issues. Such predictions cost less human effort, but they are usually biased and may lead to incorrect decisions. Thus, keeping the estimation unbiased while saving the cost becomes an important task to pursue.

In this paper, we utilize individuals' social interactions (potentially learned from social media data) to improve the unbiased sampling method. Our motivation is from the fact that people's opinions are often correlated, especially among friends in a social network, due to their social interactions in terms of the homophily and influence effects (McPherson, Smith-Lovin, and Cook 2001; Goel, Mason, and Watts 2010; Crandall et al. 2008). Such correlations are partially known in the big data era. For example, many online social media and networking sites provide publicly available social interaction data and user's sentiment data, and companies also have large amounts of data about their customers' preferences and their social interactions. Our idea is to partition individuals into different groups by utilizing the above prior knowledge, such that people within a group are likely to hold the same opinions. We can then sample very few people in each group and aggregate the sampling results together to achieve an accurate estimation. We call this the *partitioned sampling* method.

We formulate the above idea as an optimization problem. In particular, we first characterize individuals' opinions as random variables. We then specify our objective as minimizing the expected sample variance of the estimate, and define the statistical measure of pairwise *opinion similarity* as the input. Our analysis later shows that this input is enough to fully determine the solution of the optimization problem, named the *Optimal Partitioned Sampling (OPS)* problem (Section 2).

We solve the OPS problem in two steps (Section 3). First, we show that the best partition is always a *simple partition*, meaning that each group contains only one sample. Second, we use people's opinion similarities to construct a weighted graph and reduce the OPS problem to the Min-r-Partition problem. We adapt a semi-definite programming algorithm and a heuristic algorithm to solve the optimization problem. We further show that partitioned sampling using any balanced simple partition where group sizes are the same always out-performs naive sampling method, and thus balanced simple partition is always safe to use even if we only have partial or inaccurate opinion similarity information.

Next, we adapt existing opinion evolution models and propose the Voter model with Innate Opinions (VIO) based on social network interactions (Section 4). We provide an exact computation of opinion similarities in the steady state of the model, which is novel in the study of such models.

Finally, we conduct experiments on both synthetic and

*This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003, 61433014.

real-world datasets to demonstrate the effectiveness and robustness of our partitioned sampling method (Section 5).

In summary, our contributions include: (a) proposing the partitioned sampling method to improve sampling quality based on opinion similarities and formulating it as an optimization problem, (b) precisely connecting the OPS problem to the Min-r-Partition problem and providing efficient algorithms for the OPS problem, and (c) adapting an opinion evolution model and providing an exact computation of opinion similarities based on the model.

Due to space constraints, all proofs and further technical details are included in the full report (Huang, Li, and Chen 2015).

Related Work. There are many sampling methods in the literature. The most related method is stratified sampling (Bethel 1986; 1989; Chromy 1987; Cochran 2007; Kozak, Verma, and Zieliński 2007; Keskintürk and Er 2007; Ballin and Barcaroli 2013). The entire population is first stratified into homogeneous atomic strata based on individuals' profiles (e.g., age, gender, etc.), and then they may be combined to a final stratification and subsample size in each stratum is allocated to minimize sample variance. Conceptually, our partitioned sampling method is similar to stratified sampling, but there are some important differences. First, stratified sampling partitions individuals based on their profiles, which may not imply opinion similarity, while we partition individuals directly based on opinion similarity, and thus our method is more accurate and flexible. Second, the technical treatments are different. Stratified sampling treats individual opinions as fixed and unknown, and requires the (estimated) mean and standard deviation of opinions in each stratum to bootstrap the stratified sampling, while we treat individual opinions as random variables, and use pairwise opinion similarities for partitioned sampling.

Among studies on social interaction based sampling, Dasgupta, Kumar, and Sivakumar (2012) utilize social network connections to facilitate sampling. However, their method is to ask the voter being sampled to return the estimate of her friends' opinions, which changes the polling practice. In contrast, we still follow the standard polling practice and only use implicit knowledge on opinion similarities to improve sampling quality. Das et al. (2013) consider the task of estimating people's average innate opinion by removing their social interactions, which is opposite to our task — we want to utilize opinion interactions for more efficient sampling of final expressed opinions which are counted in opinion polls. Graph sampling methods (Gjoka et al. 2010; Kurant et al. 2011) aim at achieving unbiased uniform sampling on large scale networks when the full network is not available, which is orthogonal to our partitioned sampling approach and could be potentially combined.

Various opinion evolution models have been proposed in the literature (Yildiz et al. 2011; Das et al. 2013; Gionis, Terzi, and Tsaparas 2013; Li et al. 2015). Our VIO model is adapted from the voter model (Clifford and Sudbury 1973) and its extension with stubborn agents (Yildiz et al. 2011).

Graph partitioning has been well studied, and numerous problem variants and algorithms exist. In this paper, we reduce the OPS problem to the Min-r-Partition problem, which was first formulated by Sahni and Gonzalez (1976). To the best of our knowledge, there is no approximation or heuristic algorithms for Min-r-Partition. Thus, we adapt a state-of-art approximation algorithm for the dual problem (Max-r-Cut) to solve the OPS problem (Frieze and Jerrum 1997). We also propose a greedy algorithm for large graphs, which takes the idea from a heuristic algorithm for Max-r-Cut (Zhu, Lin, and Ali 2013).

2 Formulating the OPS Problem

We consider a vertex set V from a social network graph containing n vertices (or nodes) $v_1, v_2, \ldots, v_n$. Each vertex represents a person in the social network, and has a binary opinion on some topic of interest. Our task is to estimate the average opinion of all individuals in the social network with sample size budget r. Let $f : V \to \{0, 1\}$ denote the opinion function, i.e., we wish to estimate the fraction $\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(v_i)$. The *naive sampling* method simply picks r nodes uniformly at random with replacement from V to ask their opinions and takes the average of sampled opinions as the estimate, as denoted below: $\hat{f}_{naive}(V, r) = \frac{1}{r}\sum_{i=1}^{r} f(x_i)$, where x_i is the i-th sampled node.

In this paper, we propose a general sampling framework called *partitioned sampling*. Formally, we first partition the whole vertex set into several disjoint subsets (called *groups*), and then allocate subsample size of each group. We use $\mathcal{P} = \{(V_1, r_1), (V_2, r_2), \ldots, (V_K, r_K)\}$ to represent such a partition, where $V_1, V_2, \ldots, V_K$ are groups, and r_k is the subsample size of group V_k. Next, we do naive sampling inside each group V_k with its subsample size r_k. Finally, we estimate the average opinion of the population by taking a weighted average of all subsampling results, with weights proportional to group sizes: $\hat{f}_{part}(\mathcal{P}) = \sum_{k=1}^{K} \frac{|V_k|}{|V|} \cdot \hat{f}_{naive}(V_k, r_k)$. Notice that naive sampling is a special case of partitioned sampling with $\mathcal{P} = \{(V, r)\}$. One can easily verify that partitioned sampling is unbiased (Huang, Li, and Chen 2015).

Intuitively, the advantage of using partitioned sampling is that, if we partition individuals such that people likely holding the same opinions are partitioned into the same group, then we can sample very few people in each group to get an accurate estimate of the average opinion of the group, and aggregate them to get a good estimate of population mean. To implement this idea, we assume that some prior knowledge about people's opinions and their similarities is available before sampling. Based on these knowledge, our goal is to find the best partition for partitioned sampling which achieves the best sampling quality.

Our first research challenge is how to rigorously formulate the above intuition into an optimization problem. To meet this challenge, we need to answer (a) which objective function is the appropriate one for the optimization problem, and (b) which representation of the prior knowledge about people's opinions and their similarities can be used as the inputs to the optimization problem.

We first address the objective function. When all individuals' opinions $f(v_1), f(v_2), \ldots, f(v_n)$ are fixed (but unknown), the effectiveness of an unbiased randomized sampling method is measured by the standard sample variance

$\mathrm{Var}(\hat{f})$, where $\hat{f}$ is the estimate. The smaller the sample variance, the better the sampling method. When the prior statistical knowledge about people's opinions is available, effectively we treat opinions $f(v_1), f(v_2), \ldots, f(v_n)$ as random variables, and the prior knowledge is some statistics related to the joint distribution of these random variables. In this case, the best sampling method should minimize the *expected sample variance* $\mathbb{E}[\mathrm{Var}(\hat{f})]$, where the expectation is taken over the randomness from the joint distribution of people's opinions. For clarity, we use $\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$ to represent $\mathbb{E}[\mathrm{Var}(\hat{f})]$, where subscript M (standing for "model") represents the randomness from the joint distribution model of opinions, and subscript S (standing for "sampling") represents sample randomness from the sampling method. [1]

We now discuss the input to the optimization task. The full joint distribution of $f(v_1), f(v_2), \ldots, f(v_n)$ requires an exponential number of parameters and is infeasible as the input. Then notice that the objective function only involves first two moments, which suggests us to use the expectations and pairwise correlations of people's opinions as the inputs. Indeed, we find that these knowledge is good enough to fully characterize the optimization problem. However, we further discover that a weaker and more direct type of statistics would be enough to enable the optimization problem, which we formally define as pairwise opinion similarities: the *opinion similarity* σ_{ij} for nodes v_i and v_j is defined as the probability that $f(v_i)$ and $f(v_j)$ have the same values.

With the objective function and inputs settled, we are now ready to define our optimization problem:

Definition 1. *(Optimal Partitioned Sampling) Given a vertex set $V = \{v_1, v_2, \ldots, v_n\}$, sample size budget $r < n$, and opinion similarity σ_{ij} between every pair of nodes v_i and v_j, the* Optimal Partitioned Sampling *(OPS) problem is to find the optimal partition $\mathcal{P}^*$ of V, such that the partitioned sampling method using $\mathcal{P}^*$ achieves the minimum expected sample variance, i.e., $\mathcal{P}^* = \arg\min_{\mathcal{P}} \mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))]$, where $\mathcal{P}$ takes among all partitions of V with r samples.*

We remark that the OPS problem requires all pairwise opinion similarities as inputs so as to make the problem well-defined. We will address the issue of handling missing or inaccurate opinion similarities in Section 3.1, and show that partitioned sampling still has outstanding performance.

3 Solving the OPS Problem

There are two issues involved in the OPS problem: one is how to partition the vertex set V into K groups; the other is how to allocate the subsample size in each group. For simplifying the OPS problem, we first consider a special kind of partitions that pick only one sample node in each group.

Definition 2. *A* simple partition *is a partition in which the subsample size of each group is equal to one.*

Simple partitions are important not only for the simplicity but also for the superiority. We will later show in Theorem 2 that, for any non-simple partition $\mathcal{P}$, one can easily construct a simple partition based on $\mathcal{P}$ which is at least as good as $\mathcal{P}$. Thus, we focus on finding the optimal simple partition.

Our approach is constructing a weighted assistant graph G_a whose vertex set is V, where the weight of edge (v_i, v_j) is $w_{ij} = 1 - \sigma_{ij}$, and then connecting the OPS problem with a graph partitioning problem for the graph G_a. For a simple partition $\mathcal{P} = \{(V_1, 1), (V_2, 1), \ldots, (V_r, 1)\}$ of V, we use $\mathrm{Vol}_{G_a}(V_k)$ to denote the volume of the group V_k in the graph G_a, defined as $\mathrm{Vol}_{G_a}(V_k) = \sum_{v_i, v_j \in V_k} w_{ij}$. We define a cost function $g(\mathcal{P})$ to be the sum of all groups' volumes in G_a, namely, $g(\mathcal{P}) = \sum_{k=1}^{r} \mathrm{Vol}_{G_a}(V_k)$. Our major technical contribution is to show that minimizing the expected sample variance of partitioned sampling using any simple partition $\mathcal{P}$ is equivalent to minimizing the cost function $g(\mathcal{P})$, as summarized by the following theorem:

Theorem 1. *Given a vertex set V with pairwise opinion similarities $\{\sigma_{ij}\}$'s and sample size r, for any simple partition $\mathcal{P} = \{(V_1, 1), (V_2, 1), \ldots, (V_r, 1)\}$ of V,*

$$\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))] = g(\mathcal{P})/2|V|^2.$$

Thus, the optimal simple partition of V minimizes the cost function $g(\mathcal{P})$.

Proof (Sketch). We use x_k to denote the sample node selected in the k-th group V_k of the simple partition $\mathcal{P}$. The estimate of partitioned sampling with $\mathcal{P}$ can be written as $\hat{f}_{part}(\mathcal{P}) = \frac{1}{n}\sum_{k=1}^{r} n_k f(x_k)$, where $n = |V|$ and $n_k = |V_k|$. When f is fixed, since $f(x_k)$'s are independent, then

$$\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P})) = \frac{1}{n^2}\sum_{k=1}^{r} n_k^2 \cdot \mathrm{Var}_S\left[f(x_k)\right]$$

$$= \frac{1}{n^2}\sum_{k=1}^{r} n_k^2 \cdot (\mathbb{E}_S[f(x_k)^2] - \mathbb{E}_S[f(x_k)]^2).$$

We then use the fact that $f(x_k)^2 = f(x_k)$ and $\mathbb{E}_S[f(x_k)] = \sum_{v_j \in V_k} f(v_j)/n_k$, and take expectation when f is drawn from a distribution, to obtain

$$\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))] = \frac{1}{n^2}\sum_{k=1}^{r}\left((n_k - 1)\sum_{v_j \in V_k}\mathbb{E}_M[f(v_j)]\right.$$

$$\left. - \sum_{v_i, v_j \in V_k, v_i \neq v_j}\mathbb{E}_M[f(v_i)f(v_j)]\right).$$

Notice that for any two binary random variables A and B, we have $\mathbb{E}[AB] = \frac{1}{2}\left(\mathbb{P}[A = B] + \mathbb{E}[A] + \mathbb{E}[B] - 1\right)$. After applying this formula to $\mathbb{E}_M[f(v_i)f(v_j)]$ and simplifying the expression, we obtain the theorem. □

The intuition of the theorem is that, small cost function indicates small volume of each group, which implies that the nodes within each group have high opinion similarities. Theorem 1 makes precise our intuition that grouping people with similar opinions would make partitioned sampling more efficient.

Theorem 1 provides the connection between the OPS problem and the graph partitioning problem. In particular,

[1] One may propose to use the total variance $\mathrm{Var}_{M,S}(\hat{f})$ as the objective function. In the full report (Huang, Li, and Chen 2015), we show that they are equivalent for the optimization task.

Algorithm 1 Greedy Partitioning Algorithm

Require: Graph G_a with n nodes, number of groups r.

Randomly generate a node sequence of all the nodes: $x_1, x_2, \ldots, x_n$.
Let $V_1 = \ldots = V_r = \emptyset$.
repeat
  for $i \leftarrow 1$ **to** n **do**
    if $x_i \in V_j$ for some $j \in [r]$ **then** $V_j = V_j \setminus \{x_i\}$.
    end if
    $k \leftarrow \arg\min_{\ell \in [r]} \delta g_\ell(x_i, \{(V_1, 1), \ldots, (V_r, 1)\})$
    $V_k \leftarrow V_k \cup \{x_i\}$.
  end for
until a predetermined stopping condition holds.
Output: Partition $\mathcal{P} = \{(V_1, 1), \ldots, (V_r, 1)\}$.

it suggests that we can reduce the OPS problem to the following *Min-r-Partition* problem: given an undirected graph with non-negative edge weights, partition the graph into r groups such that the sum of all groups' volumes is minimized. However, Min-r-Partition is NP-hard to approximate to within any finite factor (Kann et al. 1997), and to the best of our knowledge, there is no approximation or heuristic algorithms in the literature. The good news is that Min-r-Partition and its dual problem (Max-r-Cut) are equivalent in the exact solution, and there exist both approximation and heuristic algorithms for Max-r-Cut. Frieze and Jerrum (1997) propose a semi-definite programming (SDP) algorithm which achieves $1 - 1/r + 2\ln r/r^2$ approximation ratio and is the best to date. We adopt the SDP algorithm to solve the OPS problem. The SDP partitioning algorithm including the SDP relaxation program is given in the full report (Huang, Li, and Chen 2015). The drawback of the SDP partitioning algorithm is its inefficiency. Thus, we further propose a greedy algorithm to deal with larger graphs, which takes the idea from a heuristic algorithm for Max-r-Cut (Zhu, Lin, and Ali 2013).

Given a simple partition $\mathcal{P} = \{(V_1, 1), \ldots, (V_r, 1)\}$ and an external node v_i which does not belong to V_k for any $k \in [r]$, we define $\delta g_\ell(v_i, \mathcal{P})$ to be $g(\mathcal{P}') - g(\mathcal{P})$, where $\mathcal{P}'$ is $\{(V_1, 1), \ldots, (V_\ell \cup \{v_i\}, 1), \ldots, (V_r, 1)\}$. Thus $\delta g_\ell(v_i, \mathcal{P})$ represents the increase of the cost function when the external node v_i is added to the group V_ℓ of $\mathcal{P}$. The greedy algorithm (Algorithm 1) first assigns each ungrouped node x_i to the group such that the objective function $g(\mathcal{P})$ is increased the least. After the first round of greedy assignment, the assignment procedure is repeated to further decrease the cost function, until some stopping condition holds, such as the decrease is smaller than a predetermined threshold.

The running time of one-round greedy assignment is $O(n + m)$ where m is the number of edges in G_a. In our experiment, we will show that greedy partitioning performs as well as SDP partitioning but could run on much larger graphs. Theoretically, the performance of partitioned sampling using the simple partition generated by the greedy partitioning algorithm is always at least as good as naive sampling, even using the partition generated after the first round of greedy assignment, as summarized below:

Lemma 1. *Given a vertex set V with sample size r, partitioned sampling using the simple partition $\mathcal{P}$ generated by the greedy partitioning algorithm (even after the first round) is at least as good as naive sampling. Specifically,*

$$\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))] \leq \mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{naive}(V, r))].$$

We call a partition $\mathcal{P}'$ a *refined* partition of $\mathcal{P}$, if each group of $\mathcal{P}'$ is a subset of some group of $\mathcal{P}$. Suppose we are given a partition $\mathcal{P}$ such that there exists some group which is allocated more than one sample. Then we can further partition that group by the greedy partitioning algorithm and finally obtain a refined simple partition of $\mathcal{P}$. According to Lemma 1, the refined simple partition should be at least as good as the original partition $\mathcal{P}$, summarized as below:

Theorem 2. *For any non-simple partition $\mathcal{P}$, there exists a refined simple partition $\mathcal{P}'$ of $\mathcal{P}$, which can be constructed efficiently, such that partitioned sampling using the refined simple partition $\mathcal{P}'$ is at least as good as partitioned sampling using the original partition $\mathcal{P}$. Specifically,*

$$\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}'))] \leq \mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))].$$

Theorem 2 shows the superiority of simple partitions, and justifies that it is enough for us to only optimize for partitioned sampling with simple partitions.

3.1 Dealing with Inaccurate Similarities

When accurate opinion similarities are not available, one still can use a *balanced partition* (i.e., all groups have the exact same size) to achieve as least good sampling result as naive sampling, summarized as below:

Theorem 3. *Given a vertex set V with n nodes and sample size r where n is a multiple of r, partitioned sampling using any balanced simple partition $\mathcal{P}$ is at least as good as naive sampling. That is, $\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P})) \leq \mathrm{Var}_S(\hat{f}_{naive}(V, r))$ holds for any fixed opinions $f(v_1), \ldots, f(v_n)$.*

Theorem 3 provides a safety net showing that partitioned sampling would not hurt sampling quality. Thus, we can always use the greedy algorithm with a balance partition constraint to achieve better sampling result. The result will be further improved if opinion similarities get more accurate.

Furthermore, in the experiment on the real-world dataset (Section 5), we artificially remove all the opinion similarity information (set as 0.5) between disconnected individuals, and perturb the rest opinion similarities more than 30%, to simulate the condition of missing and inaccurate similarities. The experimental result shows that the performance of the greedy algorithm with perturbed inputs is quite close to the performance of the greedy algorithm with exact inputs. This demonstrates the robustness of our greedy algorithm in the face of missing and inaccurate opinion similarity data.

Moreover, since real-world social interaction can be characterized well by opinion evolution models, we adapt a well-known opinion evolution model and give an exact computation of opinion similarity based on the model in the next section. The model essentially provides a more compact representation than pairwise similarities.

4 Opinion Evolution Model

We adapt the well-known *voter model* to describe social dynamics (Clifford and Sudbury 1973; Yildiz et al. 2011). Consider a weighted directed social graph $G = (V, A)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is the vertex set and A is the weighted adjacency matrix. Each node is associated with both an *innate opinion* and an *expressed opinion*. The innate opinion remains unchanged from external influences, while the expressed opinion could be shaped by the opinions of one's neighbors, and is the one observed by sampling. At initial time, each node v_i generates its innate opinion $f^{(0)}(v_i) \in \{0, 1\}$ from an i.i.d. Bernoulli distribution with expected value $\mu^{(0)}$. The use of i.i.d. distribution for the innate opinion is due to the lack of prior knowledge on a brand-new topic, and is also adopted in other models (Dasgupta, Kumar, and Sivakumar 2012). When $t > 0$, each node v_i updates its expressed opinion $f^{(t)}(v_i) \in \{0, 1\}$ independently according to a Poisson process with updating rate λ_i: at its Poisson arrival time t, node v_i sets $f^{(t)}(v_i)$ to its innate opinion with an *inward probability* $p_i > 0$, or with probability $(1 - p_i)A_{ij}/\sum_{k=1}^{n} A_{ik}$, adopts its out-neighbor v_j's expressed opinion $f^{(t)}(v_j)$. We call the model *Voter model with Innate Opinions (VIO)*.

The VIO model reaches a steady state if the joint distribution of all node's expressed opinions no longer changes over time.[2] We use notation $f^{(\infty)}(v_i)$ to represent the steady-state expressed opinion of node v_i, which is a random variable. We assume that opinion sampling is done in the steady state, which means that people have sufficiently communicated within the social network.

To facilitate analysis of the VIO model, we take an equivalent view of the VIO model as *coalescing random walks* on an augmented graph $\overline{G} = (V \cup V', E \cup \{e'_1, e'_2, \ldots, e'_n\})$, where $V' = \{v'_1, v'_2, \ldots, v'_n\}$ is a copy of V, E is the edge set of G and $e'_i = (v_i, v'_i)$ for all i. In this viewpoint, we have n walkers randomly wandering on $\overline{G}$ "back in time" as follows. At time t, all walkers are separately located at $v_1, v_2, \cdots, v_n$. Suppose before time t, v_i is the last node who updated its expressed opinion at time $\tau < t$, then the n walkers stay stationary on their nodes from time t until time τ "back in time". At time τ, the walker at node v_i takes a walk step: she either walks to v_i's out-neighbor $v_j \in V$ with probability $(1-p_i)A_{ij}/\sum_{k=1}^{n} A_{ik}$, or walks to $v'_i \in V'$ with probability p_i. If any walker (e.g., the walker starting from node v_i) walks to a node (e.g., v'_k) in V', then she stops her walk. In the VIO model language, this is equivalent to saying that v_i's opinion at time t is determined by v_k's innate opinion, namely $f^{(t)}(v_i) = f^{(0)}(v_k)$. If two random walkers meet at the same node in V at any time, they walk together from now on following the above rules (hence the name *coalescing*). Finally, at time $t = 0$, if the walker is still at some node $v_i \in V$, she always walks to $v'_i \in V'$.

We now define some key parameters based on the coalescing random walk model, which will be directly used for computing the opinion similarity later.

[2]The VIO model has a unique joint distribution for the final expressed opinions (Huang, Li, and Chen 2015).

Definition 3. *Let $\mathcal{I}_{ij}^{\ell}$ denote the event that two random walkers starting from v_i and v_j at time $t = \infty$ eventually meet and the first node they meet at is $v_\ell \in V$. Let Q be the $n \times n$ matrix where Q_{ij} denotes the probability that a random walker starting from v_i at time $t = \infty$ ends at $v'_j \in V'$.*

Lemma 2. *For $i, j, \ell \in [n]$, $\mathbb{P}\left[\mathcal{I}_{ij}^{\ell}\right]$ is the unique solution of the following linear equation system:*

$$\mathbb{P}\left[\mathcal{I}_{ij}^{\ell}\right] = \begin{cases} 0, & i = j \neq \ell, \\ 1, & i = j = \ell, \\ \sum_{a=1}^{n} \frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i}\mathbb{P}[\mathcal{I}_{aj}^{\ell}] & \\ \quad + \sum_{b=1}^{n} \frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j}\mathbb{P}[\mathcal{I}_{ib}^{\ell}], & i \neq j, \end{cases}$$

where $d_i = \sum_{j=1}^{n} A_{ij}$ is v_i's weighted out-degree. In addition, matrix Q is computed by

$$Q = \left(I - (I - P)\, D^{-1}A\right)^{-1} P,$$

where $P = \mathrm{diag}(p_1, \ldots, p_n)$ and $D = \mathrm{diag}(d_1, \ldots, d_n)$ are two diagonal matrices, and matrix $I - (I - P)\, D^{-1}A$ is invertible when $p_i > 0$ for all $i \in [n]$.

Our main analytical result concerning the VIO model is the following exact computation of pairwise opinion correlation, which directly leads to opinion similarity:

Lemma 3. *For any $i, j \in [n]$, opinion correlation ρ_{ij} in the steady state is equal to the probability that two coalescing random walks starting from v_i and v_j at time $t = \infty$ end at the same absorbing node in V'. Moreover, opinion correlation ρ_{ij} can be computed by*

$$\begin{aligned} \rho_{ij} &= \mathrm{Cor}_M\left(f^{(\infty)}(v_i), f^{(\infty)}(v_j)\right) \\ &= \sum_{k=1}^{n} Q_{ik}Q_{jk} + \sum_{\ell=1}^{n} \mathbb{P}\left[\mathcal{I}_{ij}^{\ell}\right]\left(1 - \sum_{k=1}^{n} Q_{\ell k}^2\right) \end{aligned}$$

where $\mathcal{I}_{ij}^{\ell}$ and Q are defined in Definition 3, and $\mathbb{P}\left[\mathcal{I}_{ij}^{\ell}\right]$ and Q are computed by Lemma 2.

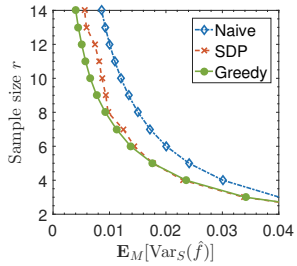
Theorem 4. *For any two nodes v_i and v_j, their opinion similarity σ_{ij} in the steady state of the VIO model is equal to:*

$$\sigma_{ij} = 1 - 2\mu^{(0)}(1 - \mu^{(0)})(1 - \rho_{ij})$$

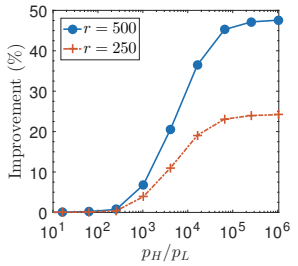
where opinion correlation ρ_{ij} is computed by Lemma 3.

Notice that for partitioning algorithms, we only need $1 - \sigma_{ij}$ as the edge weight and by the above theorem this weight value is proportional to $1 - \rho_{ij}$, which means the exact value of $\mu^{(0)}$ is irrelevant for partitioning algorithms. In the full report (Huang, Li, and Chen 2015), we will provide an efficient computation of all pairwise opinion correlations with running time $O(nmR)$ by a carefully designed iterative algorithm, where m is the number of edges of G which is commonly sparse, and R is the number of iterations. We further remark that the correlations are calculated offline based on the existing network and historical data, and thus the complexity compared to the sampling cost of telephone interview or network survey is relatively small.

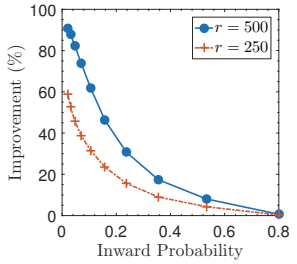
In the full report (Huang, Li, and Chen 2015), we further extend the VIO model to include (a) non-i.i.d. distributions of the innate opinions, and (b) negative edges as in the signed voter model (Li et al. 2015).



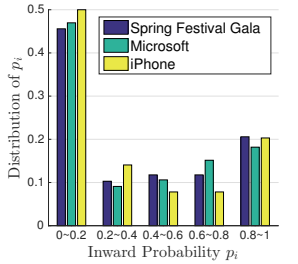
(a) Synthetic graph (100 nodes)



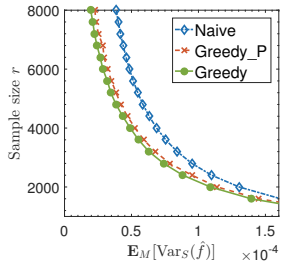
(b) Synthetic graph (10,000 nodes)



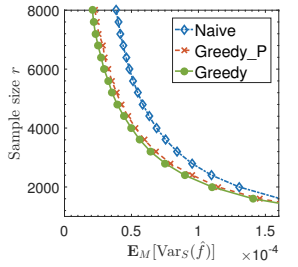
(c) Synthetic graph (10,000 nodes)



(d) Inward probability distribution



(e) Weibo graph with $\lambda_i = 1$



(f) Weibo graph with λ_i set to the number of v_i's tweets in a year

5 Experimental Evaluation

In this section, we compare the sampling quality of partitioned sampling using greedy partitioning (Greedy) and partitioned sampling using SDP partitioning[3] (SDP) against naive sampling (Naive) based on the VIO model, using both synthetic and real-world datasets. We describe major parameter settings for the experiments below, while leave the complete settings in the full report (Huang, Li, and Chen 2015) due to space constraints.

In our experiment, when the parameters of VIO model are set, the simulation is done by (a) calculating the pairwise opinion similarities by Theorem 4, (b) running the partitioning algorithms to obtain the partition candidate, and (c) computing the expected variance $\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$ by Theorem 1.

Synthetic Dataset. We use the planted partition model (Condon and Karp 2001) to generate undirected graphs, which aims at resembling the community structure in real-world social networks. Given n vertices and k latent disjoint groups, every edge (v_i, v_j) is generated with a high probability p_H if v_i and v_j are in the same latent group, otherwise with a low probability p_L.

We generate two different sizes of synthetic graphs. The small one includes 100 nodes and 20 latent groups, and p_H, p_L and λ_i are set to 0.9, 0.01 and 1, respectively. The inward probability of each node is randomly chosen from $[0, 0.01]$. Fig (a) shows that, when the sample size r is small, the performance of SDP and Greedy are similar to each other and both better than Naive. When the sample size r increases, Greedy becomes much better than Naive, and SDP starts getting worse. For the large synthetic graph with 10k nodes and 500 latent groups, SDP is no longer feasible, thus we compare the improvement of Greedy against Naive. In Fig (b), we range p_H/p_L and find that larger p_H/p_L (more apparent clustering) indicates the better performance of the partitioned sampling method. When p_H/p_L increases from 10^3 to 10^5, the improvement of expected sample variance increases rapidly. When $p_H/p_L > 10^5$, the improvement becomes saturated. This is because the number of edges which cross different latent groups are so few that it decreases rather slowly and the graph structure is almost unchanged when p_H/p_L increases further. In Fig (c), we set all nodes' inward probabilities to be equal and vary them from 0.02 to 0.8. The figure shows that the lower inward probability leads to the better performance of partitioned sampling. When the inward probability gets small, the improvement expected sample variance increases rapidly. This is because a lower inward probability means people interacting more with each other and thus their opinions are correlated more significantly. According to the above experiments, we conclude that the larger p_H/p_L and the lower inward probability make people's opinions more clustered and more correlated inside the clusters, and our partitioned sampling method works better for these cases.

Real-World Dataset. We use the micro-blog data from weibo.com (Yuan et al. 2013), which contains 100,102 users and 30,518,600 tweets within a one-year timeline from 1/1/2013 to 1/1/2014. We treat the user following relationship between two users as a directed edge (with weight 1).

We first learn the distribution of user's inward probabilities from the data. We extract a series of users' opinions on 12 specific topics (e.g., Microsoft, iPhone, etc.) by applying a keyword classifier and a sentiment analyzer (Tang et al. 2014) to the tweets. We also collect their relationships and form a subgraph for each topic. Then we use VIO model to fit the data by solving a minimization problem w.r.t. inward probabilities using gradient descent. Fig (d) shows the distribution of inward probabilities for three of the topics, namely Spring Festival Gala (68 users), Microsoft (66 users) and iPhone (59 users), and the results for other topics are similar. From these distributions, we observe that (a) over 45% inward probabilities locate in $[0, 0.2]$; (b) the probability that p_i locates in $[0.8, 1]$ is the second highest; (c) others almost uniformly locate in $[0.2, 0.8]$. This indicates that in the real world, most people tend to adopt others' opinions, which matches the intuition that people are often affected by others. We manually look up the users who locate in $[0.8, 1]$, and find that most of them are media accounts and verified users. This matches our intuition that those users always take effort to spread their own opinions on the web but rarely

[3] We use CVX package (Grant and Boyd 2014; 2008) to solve the SDP programming.

adopt others' opinions, hence they should have large inward probabilities.

Now we simulate the sampling methods on the Weibo graph. We first remove the users who do not follow anyone iteratively, and get the graph including $40{,}787$ nodes and $165{,}956$ directed edges. We generate each user's inward probability following the distribution we learned. We use two different settings for opinion updating rates: one is to set $\lambda_i = 1$ for all $i \in [n]$; the other is to set λ_i to the number of v_i's tweets in a year. The improvement of Greedy against Naive with two different updating rate settings are similar as shown in Fig (e) and (f). In particular, if we fix $\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$ to be 3.86×10^{-5}, Greedy needs 4794 samples while Naive needs 8000 samples (saving 40.1%) in Fig (e), and Greedy needs 4885 samples while Naive needs 8000 samples (saving 38.9%) in Fig (f). This indicates that partitioned sampling greatly improves the sampling quality, and the sample size saving is more apparent when the expected sample variance gets smaller (i.e., the requirement of sampling quality gets higher). Moreover, in order to test the performance of partitioned sampling with missing and inaccurate opinion similarities, we artificially remove all the opinion similarity information between disconnected nodes (set similarities as 0.5), and perturb each rest similarity σ_{ij} with a random noise e_{ij} in the range $[-0.1 - 30\% \cdot \sigma_{ij}, 0.1 + 30\% \cdot \sigma_{ij}]$ (set perturbed similarity of σ_{ij} as the median of $\{0, \sigma_{ij} + e_{ij}, 1\}$). Fig (e) and (f) show that Greedy using the above perturbed similarities (denoted as Greedy_P) is very close to Greedy, and still has a significant improvement against naive sampling.

In conclusion, the experimental results demonstrate the excellent performance of our partitioned sampling method both on synthetic and real-world datasets, even when the opinion similarities are missing or inaccurate.

References

Ballin, M., and Barcaroli, G. 2013. Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology* 39(2):369–393.

Bethel, J. W. 1986. *An optimum allocation algorithm for multivariate surveys*. US Department of Agriculture, Statistical Reporting Service, Statistical Research Division.

Bethel, J. 1989. Sample allocation in multivariate surveys. *Survey methodology* 15(1):47–57.

Chromy, J. R. 1987. Design optimization with multiple objectives. *Proceedings of the Section*.

Clifford, P., and Sudbury, A. 1973. A model for spatial conflict. *Biometrika*.

Cochran, W. G. 2007. *Sampling techniques*. John Wiley & Sons.

Condon, A., and Karp, R. M. 2001. Algorithms for graph partitioning on the planted partition model. *Random Structures & Algorithms*.

Crandall, D.; Cosley, D.; Huttenlocher, D.; Kleinberg, J.; and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In *KDD '08*.

Das, A.; Gollapudi, S.; Panigrahy, R.; and Salek, M. 2013. Debiasing social wisdom. In *KDD '13*.

Dasgupta, A.; Kumar, R.; and Sivakumar, D. 2012. Social sampling. In *KDD '12*.

Frieze, A., and Jerrum, M. 1997. Improved approximation algorithms for max k-cut and max bisection. *Algorithmica*.

Gionis, A.; Terzi, E.; and Tsaparas, P. 2013. Opinion maximization in social networks. In *SDM '13*.

Gjoka, M.; Kurant, M.; Butts, C. T.; and Markopoulou, A. 2010. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM '10*.

Goel, S.; Mason, W.; and Watts, D. J. 2010. Real and perceived attitude agreement in social networks. *Journal of Personality and Social Psychology*.

Grant, M., and Boyd, S. 2008. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*. Springer.

Grant, M., and Boyd, S. 2014. CVX: Matlab software for disciplined convex programming. http://cvxr.com/cvx.

Huang, W.; Li, L.; and Chen, W. 2015. Partitioned sampling of public opinions based on their social dynamics. *arXiv preprint arXiv:1510.05217*.

Kann, V.; Khanna, S.; Lagergren, J.; and Panconesi, A. 1997. On the hardness of approximating max k-cut and its dual. *Chicago Journal of Theoretical Computer Science*.

Keskintürk, T., and Er, S. 2007. A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics & Data Analysis* 52(1):53–67.

Kozak, M.; Verma, M. R.; and Zieliński, A. 2007. Modern approach to optimum stratification: Review and perspectives. *Statistics in Transition* 8(2):223–250.

Kurant, M.; Gjoka, M.; Butts, C. T.; and Markopoulou, A. 2011. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In *SIGMETRICS '11*.

Li, Y.; Chen, W.; Wang, Y.; and Zhang, Z. 2015. Voter model on signed social networks. *Internet Mathematics*.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*.

Sahni, S., and Gonzalez, T. 1976. P-complete approximation problems. *Journal of the ACM (JACM)*.

Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; and Qin, B. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL '14*.

Yildiz, E.; Acemoglu, D.; Ozdaglar, A. E.; Saberi, A.; and Scaglione, A. 2011. Discrete opinion dynamics with stubborn agents. *Available at SSRN 1744113*.

Yuan, N. J.; Zhang, F.; Lian, D.; Zheng, K.; Yu, S.; and Xie, X. 2013. We know how you live: Exploring the spectrum of urban lifestyles. In *COSN '13*.

Zhu, W.; Lin, G.; and Ali, M. 2013. Max-k-cut by the discrete dynamic convexized method. *INFORMS Journal on Computing*.