Partitioned Sampling of Public Opinions Based on Their Social Evolution

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

Liang Li
Microsoft Research Asia
Beijing, China
liangl@microsoft.com

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

ABSTRACT

Public opinion polling is typically done by random sampling from the entire population, treating the opinions of individuals as independent. In the real world, individuals' opinions are often correlated, especially among friends in a social network, due to the effect of both homophily and social influence. In this paper, we propose a partitioned sampling method, utilizing the correlations between individuals' opinions to improve the sampling quality. In particular, we propose an adaptation of an opinion evolution model in social networks, and formulate an optimization problem based on this model as finding the optimal partition for the partitioned sampling method to minimize the expected sample variance of the estimated result. For the opinion evolution model, we develop an efficient and exact computation of opinion correlations between every pair of nodes in the social network. For the optimization task, we show that when the population size is large enough, the complete partition which contains only one sample in each component is always better, and utilize the correlation computation result obtained earlier to reduce finding optimal complete partition to a well-studied Max-r-Cut problem. We adopt the semidefinite programming algorithm for Max-r-Cut to solve our optimization problem, and further develop a greedy heuristic algorithm to improve the efficiency. We use both synthetic and real-world datasets to demonstrate that our partitioned sampling method results in significant improvement in sampling quality.

Categories and Subject Descriptors

G.3 [**Discrete Mathematics**]: Probability and Statistics—*Markov processes*; J.4 [**Computer Applications**]: Social and Behavior Science

General Terms

Design, Theory, Experimentation

Keywords

sampling, opinion evolution dynamics, social networks

1. INTRODUCTION

Public opinion polling is a main tool for governments, organizations and companies to gather information about public sentiments on the policies, strategies, products etc., which are important in organizational decision making. Opinion polling needs to be accurate and unbiased, and thus it is usually done by randomly sampling a large enough number of individuals from the entire population, but this is a costly effort. Therefore, saving the cost on unbiased random sampling while keeping the same sampling quality is an important task to pursue.

In this paper, we utilize individuals' social interactions to improve the random sampling method. Our motivation is that people's opinions are often correlated, especially among friends in the social network, due to their social interactions in terms of the homophily and influence effects [5, 11, 16]. In the era of big data, these social interactions and correlations are partially known. For example, many online social media and social networking sites provide public available social interaction 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 such partial knowledge, such that people within the same group are likely to hold similar opinions on a topic of interest. We can then sample each group separately and aggregate the samplings together to achieve the more efficient sampling result. We call this *partitioned sampling* method.

More specifically, we first propose an opinion evolution model in social networks called *Voter model with Innate Opinions (VIO)*, adapted from the voter model often used to characterize opinion evolution dynamics. We then precisely state our problem as an optimization problem called *Optimal Partitioned Sampling (OPS)*: find the optimal partition of nodes in a social network and the sample size allocation to each component, with the goal of minimizing expected sample variance based on the VIO model (Section 2).

†Institute for Interdisciplinary Information Sciences

*This work was done when the first author was visiting Microsoft Research Asia as a research intern. 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.

We provide a precise analysis of the VIO model, including an efficient and exact computation of the correlations between the final opinions of every pair of nodes in steady state of the VIO model (Section 3). We then study the OPS problem, reaching two important conclusions (Section 4). First, we show that when population size is large enough, the best partition is always the complete partition, meaning that each component only contains one sample. Second, we use computed correlations to build a weighted graph and reduce OPS problem to a weighted graph partitioning problem, which is a special case of the well-studied Max-r-Cut problem. We then adopt the semidefinite programming algorithm for Max-r-Cut to solve OPS, and further propose an efficient greedy partitioning algorithm to work on larger graphs.

Finally, we conduct experiments on both synthetic and real-world datasets to demonstrate that our partitioned sampling method indeed improves sampling quality over traditional naive sampling method, which translates into significant cost savings if we maintain sampling quality at the same level (Section 5).

In summary, our contribution includes: (a) proposing the partitioned sampling method and formulating it as an optimization problem to improve sampling quality based on social interactions and correlations of opinions; (b) adapting an opinion evolution model and providing exact solutions for computing the key quantities of the model; and (c) precisely connecting the optimal partitioned sampling problem to the Max-r-Cut problem and providing efficient algorithms for the partitioned sampling under the opinion evolution model. We remark that our technical result on OPS problem is not constrained to our VIO model, and has wider applicability as explained after the main Theorem 3.

Related work. To the best of our knowledge, there is no other technical work on partitioned sampling. Among studies on population sampling, Dasgupta et al. [7] also utilize social network connections to facilitate sampling. However, their method is to explicitly ask the subject being sampled to return additional information about their friends' opinions and the number of their friend's friends, which requires changing the polling practice. Our partitioned sampling method, on the other hand, still follows the standard polling practice and only uses implicit knowledge on opinion correlations to improve sampling quality. These two ideas are orthogonal and could be potentially combined together. Das et al. [6] study the task of removing the correlations among individual's opinions due to their social interactions to obtain the average *original innate opinion*. Their task is to utilize the wisdom of the crowd for extracting the latent independent opinions of individuals. Our task is exactly the opposite — we want to utilize opinion interactions and correlations for more efficient sampling of *final expressed opinions*, which are what being counted for in opinion polling.

Various opinion evolution models have been proposed in the literature (e.g. [6, 10, 14, 19]). Our VIO model and its analysis are adapted from the voter model [2] and its extension with stubborn agents [19]. The models in [6, 10] also distinguish between innate opinions and expressed opinions, however, their models are deterministic, and thus their analyses do not apply to our stochastic analysis on the correlations between final expressed opinions.

Graph partitioning has been well studied, and numerous problem variants and algorithms exist. In this paper, we reduce the OPS problem to the Max-r-Cut problem, which is the problem of partitioning the graph in to r components and maximizing the sum of edge weights on the cut, and we adopt the semidefinite programming algorithm proposed in [8].

Algorithm 2.1 Partitioned Sampling

Require: Partition $\mathcal{P} = \{(V_1, r_1), (V_2, r_2), \ldots, (V_K, r_K)\}$
1: **for** $k \leftarrow 1$ **to** K **do**
2: Do naive sampling in V_k, **return** $\hat{f}_{naive}(V_k, r_k)$.
3: **end for**
4: **Output:** $\hat{f}_{part}(\mathcal{P}) = \sum_{k=1}^{K} \frac{|V_k|}{|V|} \cdot \hat{f}_{naive}(V_k, r_k)$.

2. OPINION EVOLUTION MODEL AND PARTITIONED SAMPLING PROBLEM

We consider a weighted directed social graph $G = (V, A)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is the vertex set containing n vertices (or nodes) and A is the weighted adjacency matrix. An edge (v_i, v_j) exists if the entry $A_{ij} > 0$. The nodes represent the individuals in the social network, and edge (v_i, v_j) represents the opinion influence relationship from v_j to v_i.

Each individual in the social network has a binary opinion on some topic of interest. Our task is to do efficient sampling with sample size budget r for estimating the average opinion of the entire population (all people in the social network). The most intuitive method is called *naive sampling*, which picks r people randomly without replacement from V to ask their opinions and takes the average of sampled opinions as the estimate, and we denote it as $\hat{f}_{naive}(V, r)$.

In this paper, we propose a general sampling framework called *partitioned sampling* (Algorithm 2.1). We then utilize the partial knowledge about people's social interactions and opinion correlations to find the best partition for partitioned sampling. To be more specific, we first partition the whole graph into several components, and allocate the subsample size in each component. We use notation $\mathcal{P} = \{(V_1, r_1), (V_2, r_2), \ldots, (V_K, r_K)\}$ to represent a partition, where V_1, V_2, $\ldots$, V_K are disjoint sets with $\cup_{i=1}^{K} V_i = V$ and r_i is the subsample size allocated in V_i with $\sum_{i=1}^{K} r_i = r$. Next we do naive sampling inside each component V_i with sample size r_i. Finally, we estimate the average opinion of the entire population by taking the weighted average of all subsampling results, where the weight is proportional to the size of the component. We use $\hat{f}_{part}(\mathcal{P})$ to represent the final estimate of partitioned sampling using partition $\mathcal{P}$. Notice that naive sampling is a special case of partitioned sampling, when all vertices in V are partitioned into the unique component $\mathcal{P} = \{(V, r)\}$. It is easy to verify that partitioned sampling is unbiased.

PROPOSITION 1. *Partitioned sampling given in Algorithm 2.1 is unbiased.*

Intuitively, the advantage of using partitioned sampling is that, if we can partition individuals such that people likely holding the same opinions are partitioned into the same component, then we can sample very few people in each component to get an accurate estimate of the average opinion of the component and then aggregate them to get a good estimate of the average opinion of the entire population.

We make the above idea precise in our paper by (a) explicitly modeling the public opinion evolution dynamics among individuals in a social network based on their opinion influence relationship in the network, and (b) formulating the partitioned sampling as an optimization problem to minimize the expected sample variance based on the evolution model.

In the social network, one's opinion is often affected by her friends, leading to opinion clustering in the network. Voter model [2] is a popular one used to describe such opinion dynamics, and various extensions exist, such as the model in [19] that allows stubborn agents who do not change their own opinions. In this paper, we further extend the model in [19] to allow a person to either keep her own innate opinion or adopt a friend's opinion (similar in concepts as the model in [6, 10]), and also allow different individuals to update their opinions with distinct rates.

More specifically, each node in the social graph is associated with both an *innate* opinion and an *expressed* opinion for any given topic. 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. We call this adapted model *Voter model with Innate Opinions (VIO)*, and describe its technical detail below.

For each node v_i in the graph, let $f_i(t) \in \{0, 1\}$ denote its expressed opinion at time t, for $t \geq 0$. At initial time $t = 0$, each node v_i generates its innate opinion $f_i(0)$ from an i.i.d. distribution with mean μ. The use of i.i.d. distribution for the initial opinion is due to the lack of prior knowledge when the initial opinion for a brand-new topic is formed, and has been used in other models before (e.g. [7]). Each node v_i updates its expressed opinion according to the Poisson process with rate λ_i independently. In particular, node v_i, at each Poisson arrival time t, sets its expressed opinion $f_i(t)$ to be its own innate opinion $f_i(0)$ with an *inward probability* p_i, or with probability $1 - p_i$, node v_i randomly selects one of its out-neighbors v_j with probability proportional to the weight of the edge (v_i, v_j) (i.e. with probability $(1 - p_i)A_{ij}/\sum_{s=1}^{n} A_{is}$ where A is the weighted adjacency matrix) and sets its expressed opinion $f_i(t)$ to be $f_j(t)$. Our model degenerates to the original voter model when all the inward probabilities equals to zero and all the Poisson rates are identical, and to the model with stubborn agents [19] when all stubborn agents have $p_i = 1$ and other agents have $p_i = 0$, meanwhile, all agents have $\lambda_i = 1$. Thus the inward probability p_i represents the inward tendency (or stubbornness) of node v_i. In summary, our VIO model is parametrized by the weighted adjacency matrix A, the inward probabilities $p_1, p_2, \ldots, p_n$, the Poisson rates $\lambda_1, \lambda_2, \ldots, \lambda_n$, and the mean of innate opinion μ.

The VIO model reaches a steady state if the joint opinion distribution of $f_1(t)$, $f_2(t)$, $\ldots$, $f_n(t)$ no longer change over time. We use random variable f_i to represent the steady-state expressed opinion of node v_i. We will show in Section 3 that when inward probability $p_i > 0$ for all i, the steady state is unique. We assume that opinion sampling is done when the system reaches the steady state, with the target of estimating the average expressed opinion of the entire population $\bar{f} = \sum_{i=1}^{n} f_i/n$. This means that people have sufficiently communicated within the social network about their opinions on the topic of interest before the sampling is done, which is a reasonable assumption.

Given the VIO model, our goal is to find the best partition of all nodes to achieve the most effective sampling result. By convention, if the opinions of nodes $f_1, f_2, \ldots, f_n$ are fixed, the effectiveness of a random sampling method is measured by the sample variance $\mathrm{Var}(\hat{f})$, where $\hat{f}$ is the estimated result based on the sampling method and the variance is taken from the randomness of the sampling method. For example, for naive sampling, the variance is taken from the randomness of sample selection. Of course, we assume that the estimate is unbiased, that is, $\mathbb{E}[\hat{f}] = \bar{f}$. The best sampling method should minimize the sample variance $\mathrm{Var}(\hat{f})$. Similarly, when the opinions $f_1, f_2, \ldots, f_n$ are random variables due to the evolution of opinion, and we have the partial knowledge about the evolution (such as the parameters of the VIO model but not the actual randomness taken during the evolution), the best sampling method should minimize the *expected sample variance* $\mathbb{E}[\mathrm{Var}(\hat{f})]$, where the expectation is taken over the randomness in the opinion evolution model. To clarify the source of randomness, henceforth we use subscript M to represent model randomness from the evolution model, and subscript S to represent sample randomness from the sampling method, and thus $\mathbb{E}[\mathrm{Var}(\hat{f})]$ is clarified as $\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$.

With the objective function clearly defined as above, we are now ready to formulate our optimization problem of finding the best partition for the partitioned sampling:

Definition 1. *(Optimal Partitioned Sampling) Suppose that a social network $G = (V, A)$ follows the VIO opinion evolution model with inward probabilities $p_1, p_2, \ldots, p_n$ and Poisson rates $\lambda_1, \lambda_2, \ldots, \lambda_n$. Given the sample size budget r, the* Optimal Partitioned Sampling *problem is to find the optimal partition $\mathcal{P}$ of V with the corresponding sample size allocation, such that when we use partitioned sampling method as given in Algorithm 2.1 with the above partition $\mathcal{P}$ as input, the expected sample variance $\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))]$ is minimized.*

Note that in the OPS definition, we do not specify the mean of innate opinion μ as an input. We will show below that μ is not needed for the optimization task.

3. VIO MODEL ANALYSIS

In this section, we provide the analysis of the evolution model VIO, in preparation for the optimization task in the next section. For the reason to be made clear in the next section, the key quantities we want to obtain from the VIO model are the correlations between every two individuals' expressed opinions in steady state. We use notation $\mathrm{Cor}_M(f_i, f_j)$ to represent the correlation between v_i's and v_j's expressed opinions in steady state. From now on, we only study the VIO model with $p_i > 0$ for all $i \in [n]$, that is, individuals always leave some chance for their innate opinions.

3.1 Random-Walk Based Analysis

To facilitate the analysis, we construct an augmented graph $\overline{G}$ (Figure 1) from the original social graph G as follows (similar to the construction in [10]). Based on the social graph $G = (V, A)$, we add a new vertex set $V' = \{v_i'\}_{i=1}^{n}$, which is a copy of the vertex set V. Each vertex $v_i \in V$ connects to its corresponding vertex $v_i' \in V'$ with a directed edge $e_i' = (v_i, v_i')$. Thus the augmented

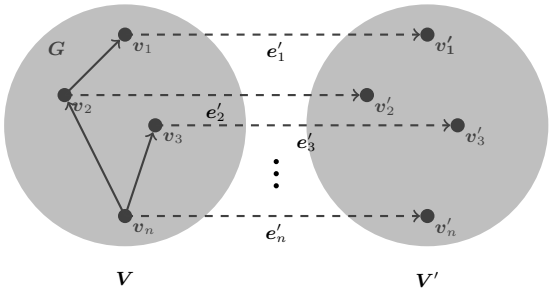


Figure 1: The augmented graph $\overline{G}$: The left gray circle represents the original social graph $G = (V, A)$; The right gray circle represents the added vertex set V', the copy of V. The dashed lines represent the new directed edges connecting the corresponding nodes between V and V'.

graph $\overline{G} = (V \cup V', E \cup \{e'_1, e'_2, \ldots, e'_n\})$ is established where E is the edge set of G.

The voter model and its variants are often analyzed through the equivalent *coalescing random walks* (e.g. [4, 15, 19]). We now specify the coalescing random walks for the VIO model on the augmented graph $\overline{G}$. For each node v_i, let $N_i = \{v_j : A_{ij} > 0\}$ be the set of its out-neighbors and $d_i = \sum_{v_j \in N_i} A_{ij}$ be its (weighted) out-degree. We consider random walkers walking on the graph $\overline{G}$, but "back in time". To know the state of v_i at time t, we start a random walker on node v_i at time t, who walks "back in time" until she reaches a node in $v'_s \in V'$, and then the innate opinion of v_s, $f_s(0)$, is the opinion of v_i at time t. The random walk goes as follows. Let the last Poisson arrival of node v_i before t came at time $\tau < t$. The random walker started on v_i at time t stays on v_i "back in time" until time τ, at which time she either walks to v_i's out-neighbor $v_j \in V$ with probability $(1 - p_i)A_{ij}/d_i$, or walks to $v'_i \in V'$ with probability p_i. Notice that this random walk step is exactly like one step of node v_i when it decides which opinion to adopt at time τ in the VIO model. If she reaches node $v_j \in V$ at time τ, she continues the walk "back in time" in the same manner, but this time she uses the previous Poisson arrival time $\tau' < \tau$ of node v_j to determine when she starts her next walk step from v_j. At any time, if the walker reaches a node $v'_s \in V'$, then the walk stops (is absorbed), and v_i's opinion at time t is determined to be $f_s(0)$. If two random walkers meet at the same node in V at any time, then they will walk together from now on following the above rule (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 $v'_i \in V'$. It is straightforward to verify that the coalescing random walk model is equivalent to the VIO model, in that for every fixed innate opinions $f_1(0), f_2(0), \ldots, f_n(0)$, the joint distribution of $f_1(t), f_2(t), \ldots, f_n(t)$ of the VIO model is the same as the joint distribution of n walkers' final opinions when they reach V', if they start the walks at nodes $v_1, v_2, \ldots, v_n$ respectively at time t.

Note that when we study the steady state of the VIO model, the time t tends to infinity, and since $p_i > 0$ for all i, all random walkers reach V' before time 0 with probability 1, and thus the special random walk rule for $t = 0$ is not essential. Thus, we also say that the steady state behavior is when all random walkers start their random walks at time $t = \infty$. With the coalescing random walk model, we can show that there is a unique steady state for the VIO model:

LEMMA 1. *When $p_i > 0$ for all $i \in [n]$, the VIO model has a unique joint distribution for the final expressed opinions in steady state.*

Since the steady state is unique and reachable, thus we can analyze the stochastic quantities of the steady state. First, we show that the expectation of any node v_i's final expressed opinion is equal to the mean of innate opinion.

LEMMA 2. *The expected expressed opinion of each node in steady state is equal to the mean of innate opinion, that is $\mathbb{E}_M[f_i] = \mu$ for all $i \in [n]$.*

We then focus on the study of the opinion correlations between each pair of nodes. To do so, we provide some key definitions related to the coalescing random walk model, together with their analysis below.

DEFINITION 2. *Let $\mathcal{I}^l_{ij}$ denote the event that two random walks starting from v_i and v_j at time $t = \infty$ eventually meet and the first node they meet at is $v_l \in V$. Let Q be the $n \times n$ matrix where Q_{ij} denotes the probability that a random walker starting from node v_i at time $t = \infty$ ends at $v'_j \in V'$.*

The following lemma provides the exact computation of $\mathbb{P}\left[\mathcal{I}^l_{ij}\right]$ and Q.

LEMMA 3. *$\mathbb{P}\left[\mathcal{I}^l_{ij}\right]$, $i, j, l \in [n]$ is the unique solution of the following linear equation system:*

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

Q is computed by

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

where $P = \mathrm{diag}(p_1, p_2, \ldots, p_n)$ and $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ are two diagonal matrices, and matrix $I - (I - P)\, D^{-1} A$ is invertible when $p_i > 0$ for all $i \in [n]$.

With $\mathbb{P}\left[\mathcal{I}^l_{ij}\right]$ and Q computed, we can obtain the correlation between any two expressed opinions. The following theorem provides our main analytical result concerning the VIO model.

THEOREM 1. *For any $i, j \in [n]$, correlation $\mathrm{Cor}_M(f_i, f_j)$ 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, $\mathrm{Cor}_M(f_i, f_j)$ can be computed by*

$$\mathrm{Cor}_M(f_i, f_j) = \sum_{l=1}^n \mathbb{P}\left[\mathcal{I}^l_{ij}\right]\left(1 - \sum_{k=1}^n Q^2_{lk}\right) + \sum_{k=1}^n Q_{ik}Q_{jk},$$

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

Note that the correlation between two expressed opinions in steady state only depends on the social network structure and individual's inward tendency, but it does not depend on the mean of the innate opinion μ. This matches our intuition that people's opinion similarity (correlation) is generated due to their social interactions, and it also facilitates our optimization in Section 4.

3.2 Efficient Correlation Computation

Naive computation directly using Theorem 1 and Lemma 3 by solving the linear equation system for $\{\mathbb{P}\left[\mathcal{I}_{ij}^l\right]\}$ would have a running time of $O(n^7)$ (See in proof of Lemma 3). Instead, we can do iterative computation on $\{\mathbb{P}\left[\mathcal{I}_{ij}^l\right]\}$ to reduce the running time to $O(n^4R)$, where R is the number of iterations. We now further improve the running time to $O(n^3R)$ by a more carefully designed iterative computation method.

We use notation C to denote the correlation matrix whose (i,j) entry is $\mathrm{Cor}_M(f_i,f_j)$, notation C' to denote the matrix whose (i,j) entry is $\sum_{l=1}^n \mathbb{P}\left[\mathcal{I}_{ij}^l\right]\left(1-\sum_{k=1}^n Q_{lk}^2\right)$. Thus, we can use the matrix form to represent the correlation computing equations as

$$C = C' + QQ^{\mathrm{T}}. \tag{1}$$

When $i \neq j$,

$$\begin{aligned}
C'_{ij} =& \sum_{l=1}^n \mathbb{P}\left[\mathcal{I}_{ij}^l\right]\left(1-\sum_{k=1}^n Q_{lk}^2\right)\\
=& \sum_{a=1}^n \frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i}\sum_{l=1}^n \mathbb{P}[\mathcal{I}_{aj}^l]\left(1-\sum_{k=1}^n Q_{lk}^2\right)\\
&+\sum_{b=1}^n \frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j}\sum_{l=1}^n \mathbb{P}[\mathcal{I}_{ib}^l]\left(1-\sum_{k=1}^n Q_{lk}^2\right)\\
=& \sum_{a=1}^n \frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i}C'_{aj} + \sum_{b=1}^n \frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j}C'_{ib}.
\end{aligned}$$

We use notation B to denote the matrix whose (i,j) entry is $\frac{\lambda_i(1-p_i)}{(\lambda_i+\lambda_j)d_i}$. Notice that C' is symmetric, thus we can write C'_{ij} in matrix form as

$$C' = B\,.*\,AC' + (B\,.*\,AC')^{\mathrm{T}}. \tag{2}$$

where $.*$ is the operator for element-wise multiplication which is used in MATLAB. The above equation holds except for the diagonal entries of C', and the i-th diagonal entry C'_{ii} equals to $1-(QQ^{\mathrm{T}})_{ii}$ for all $i \in [n]$. Therefore, we can use the iterative procedure to compute the matrix C' (i.e., in each iterative step, compute the new C' by Equation (2) and set new C''s diagonal entries to be $\mathrm{diag}(I-QQ^{\mathrm{T}})$), and finally obtain the correlation matrix C by Equation (1).

The running time of the above method with one iteration is $O(n^3)$, which depends on the running time of matrix multiplication. Simulation results show that C' can be accurately computed through a small number of iterations.

4. SOLVING OPS PROBLEM

In this section, we turn to address the OPS problem based on the VIO model. A partition is called *complete* if there is only one sample in each partition component. Our first observation is that we should always seek the complete partition in order to achieve better performance (when population size n is large enough). A partition is called *balanced* if the size difference of any two components is no more than one.

LEMMA 4. *Given a graph with $n \geq 2$ vertices, partitioned sampling using any balanced complete partition $\mathcal{P}$ of the graph, is better than naive sampling from the graph (after ignoring an $o(1)$ term). Specifically,*

$$\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P})) < \mathrm{Var}_S(\hat{f}_{naive}) + \frac{3}{2n}.$$

We call a partition the *refined partition* of $\mathcal{P}$, if its each component is the subset of some component of $\mathcal{P}$. Suppose we have a partition $\mathcal{P}$ such that there exists a component containing at least two samples. If we further partition that component by a complete partition, we can obtain a refined partition of $\mathcal{P}$, and according to the above lemma, the refined partition would be better than the original partition $\mathcal{P}$. This leads us to seek the complete partition for achieving better performance.

THEOREM 2. *Given a graph G and the VIO model on G, for any partition $\mathcal{P}$, partitioned sampling using the refined complete partition $\mathcal{P}'$ of $\mathcal{P}$ is better than partitioned sampling using the original partition $\mathcal{P}$ (after ignoring an $o(1)$ term). Specifically,*

$$\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}'))] < \mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))] + \frac{3}{2n}.$$

In the above theorem, we show that partitioned sampling using the refined complete partition of any incomplete partition is better than partitioned sampling using the original incomplete partition when n is large enough. Henceforth, we focus on finding the best complete partition. For convenience, we also use $\mathcal{P} = \{V_1, V_2, \ldots, V_r\}$ to denote the complete partition $\mathcal{P} = \{(V_1,1),(V_2,1),\ldots,(V_r,1)\}$ in the following contents.

We construct an assistant graph G_a whose vertex set is V and weight w_{ij} between v_i and v_j is $1-\mathrm{Cor}_M(f_i,f_j)$. Given a partition $\mathcal{P} = \{V_1, V_2, \ldots, V_r\}$ of V, we denote the volume of component V_k on G_a as $\mathrm{Vol}_{G_a}(V_k) = \sum_{v_i,v_j \in V_k} w_{ij}$. The *cost function* $g_r(\mathcal{P})$ is defined to be the sum of all components' volumes on G_a, namely, $g_r(\mathcal{P}) = \sum_{k=1}^r \mathrm{Vol}_{G_a}(V_k)$. Our major technical contribution is to show that minimizing the expected sample variance of complete partitioned sampling is equivalent to minimizing the sum of all components' volumes on G_a, as summarized in the following theorem:

THEOREM 3. *For any complete partition $\mathcal{P}$,*

$$\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))] = \frac{\mu(1-\mu)}{n^2} g_r(\mathcal{P}),$$

where μ is the mean of innate opinion. Thus, the best complete partition minimizes the cost function.

The above theorem is the key to bridge between our objection function on expected sample variance and the graph partition method.

Intuitively, small cost function indicates small volumes of all the components, and thus the edge weight between each pair of nodes in the same component is small, which means their correlations are high. This means that for partitions with small cost functions, the nodes in the same component tend to be strongly correlated and thus the nodes we sample can effectively represent the opinions of their respective components. Therefore, Theorem 3 makes precise our intuition that grouping people with similar opinion tendencies together and sample each group separately would make sampling more efficient. We further remark that Theorem 3 actually holds for any joint distribution of $f_1, f_2, \ldots, f_n$ when

Algorithm 4.1 SDP Partitioning Algorithm

Require: Graph G_a with n nodes, partition size r.

1: Solve the following SDP problem and compute the Cholesky decomposition of Y. Let $y_1, y_2, \ldots, y_n$ be the resulting vectors.

$$\text{Maximize} \quad \sum_{i,j} [1 - \text{Cor}(f_i, f_j)] (1 - Y_{ij}) \qquad \text{(SDP)}$$

$$\text{Subject to} \quad (a)\ Y_{ii} = 1, \forall i, (b)\ Y_{ij} \geq -\frac{1}{r-1}, \forall i \neq j,$$

$$(c)\ Y \succeq 0, (d)\ Y \text{ is symmetric.}$$

2: Choose r random vectors $z_1, z_2, \ldots, z_r$ from $\mathbb{R}^n$.

3: Partition V into r components $V_1, \ldots, V_r$ according to which of $z_1, z_2, \ldots, z_r$ is closest to each y_k.[1]

4: **Output:** $\mathcal{P} = \{(V_1, 1), \ldots, (V_r, 1)\}$.

$\mathbb{E}[f_1], \mathbb{E}[f_2], \ldots, \mathbb{E}[f_n]$ are identical, not limiting to distributions generated by the VIO model. Therefore, it can be potentially applied to a wider range of the partitioned sampling situations.

Theorem 3 suggests that we can reduce the OPS problem to the following *Min-r-Volume* problem: given an undirected graph with edge weights, partition the graph into r components such that the sum of all components' volumes is minimized. However, the Min-r-Volume problem contains r-Coloring problem as a special case, which has minimum volume of zero if any only if the graph is r-colorable. This leads to the following strong inapproximability result:

LEMMA 5. *The Min-r-Volume problem is NP-hard to be approximated within any finite factor.*

Note that the above hardness does not directly imply the hardness of OPS, since the assistant graph G_a generated by the VIO model is of particular form with edge weights $1 - \text{Cor}_M(f_i, f_j)$. We leave the hardness of OPS as an open question, and next we use the dual problem of Min-r-Volume to help solving the OPS problem.

The dual problem of Min-r-Volume is the following Max-r-Cut problem: given an undirected weighted graph, partition the graph into r components such that the total edge weight of the cut (set of edges crossing different components) is maximized. It is clear that the two problems are equivalent in terms of exact solutions, but they are different in terms of approximability. In particular, Frieze and Jerrum [8] show that for the Max-r-Cut problem, a semi-definite programming (SDP) based partition achieves $1 - 1/r + 2\ln r/r^2$ approximation ratio.

We adopt the SDP algorithm to solve the OPS problem. The SDP algorithm including the SDP relaxation program is given in Algorithm 4.1 (the original IP formulation is given in Appendix B).

The drawback of the SDP partitioning algorithm is that it is rather slow. Thus we further propose a heuristic greedy algorithm to solve the Min-r-Volume problem, which can be applied to large graphs. Given disjoint node sets $V_1, \ldots, V_r$ and an external node v_i not in any of these sets, we define $\delta g_r(v_i, V_k)$ to be $g_r(V_1, \ldots, V_k \cup \{v_i\}, \ldots, V_r) - g_r(V_1, \ldots, V_k, \ldots, V_r)$, which is the increase of the cost function when the external node v_i is added to the component V_k. The basic idea of our greedy algorithm (Algorithm 4.2) is to assign each node to the component such that the objective function $g_r(\mathcal{P})$ is increased the least. After the first round of greedy assignment, we repeat the greedy assignment procedure to further decrease the cost function, until some stopping condition holds, such as the relative decrease is smaller than a predetermined threshold.

[1]If the partitioning result is less than r components, just reselect r new random vectors from $\mathbb{R}^n$ and repeat the step again.

Algorithm 4.2 Greedy Partitioning Algorithm

Require: Graph G_a with n nodes, partition size r.

1: Generate a random permutation of the integers from 1 to n inclusive: $s_1, s_2, \ldots, s_n$.
2: Let $V_1 = \ldots = V_r = \emptyset$.
3: **repeat**
4: **for** $i \leftarrow 1$ **to** n **do**
5: **if** $v_{s_i} \in V_j$ for some $j \in [r]$ **then** $V_j = V_j \setminus \{v_{s_i}\}$.
6: **end if**
7: Compute $k = \arg\min_{l \in [r]} \delta g_r(v_{s_i}, V_l)$.
8: Let $V_k = V_k \cup \{v_{s_i}\}$.
9: **end for**
10: Let $\mathcal{P} = \{(V_1, 1), \ldots, (V_r, 1)\}$.
11: **until** a predetermined stopping condition holds
12: **Output:** $\mathcal{P}$.

Even though the greedy algorithm is a heuristic, the following lemma shows that it always performs better than the naive sampling algorithm (when ignoring an $o(1)$ term for large graphs), even using the partition after the first round of greedy assignment.

LEMMA 6. *Let $\mathcal{P}$ be the partition produced by greedy partitioning algorithm (Algorithm 4.2) after the first iteration of all nodes. Then*

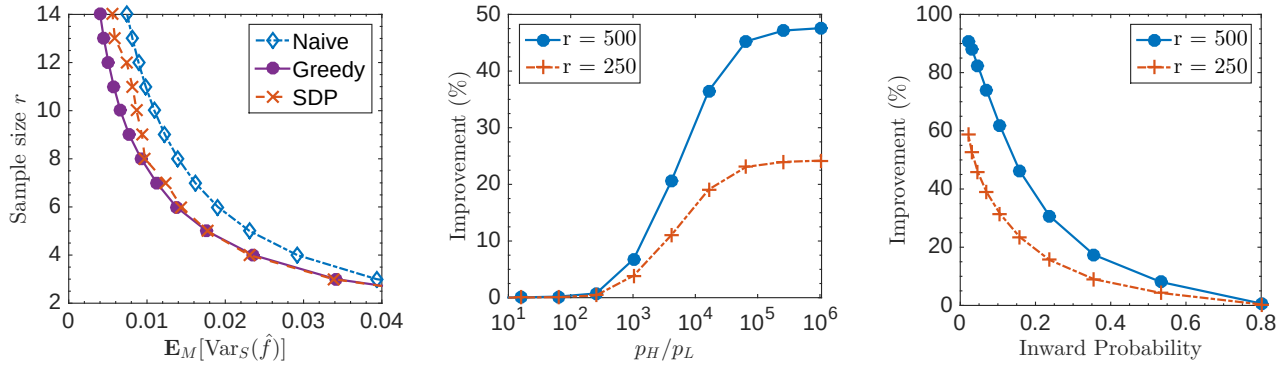
$$\mathbb{E}_M[\text{Var}_S(\hat{f}_{part}(\mathcal{P}))] < \mathbb{E}_M[\text{Var}_S(\hat{f}_{naive})] + \frac{1}{4n}.$$

The running time of one-round greedy partitioning is $O(n + |E_{G_a}|)$ where n is $|V|$ and E_{G_a} is the edge set of G_a. If G_a is strongly connected, then $|E_{G_a}| = n(n-1)/2$, thus the complexity of greedy partitioning is $O(n^2)$. In our experiment, we will show that greedy partitioning with a reasonable stopping condition performs close to SDP partitioning but could run on much larger graphs.

5. EXPERIMENTAL EVALUATION

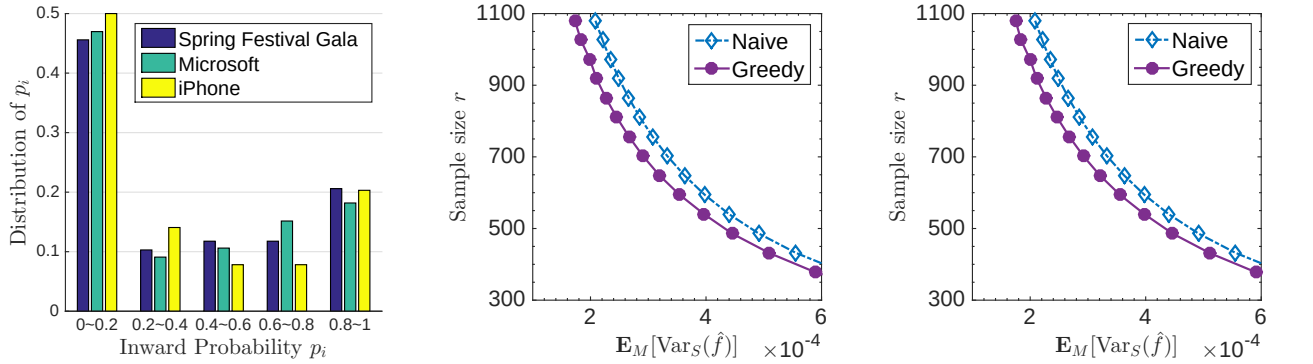
In this section, we present results of our experimental evaluations of the sampling quality of our partitioning algorithms proposed in Section 4 compared against naive sampling, using both synthetic and real-world datasets. To be specific, we first describe the generation of our synthetic graphs, and show the comparison of various sampling methods and how graph structure and inward tendency affect the performance of the sampling methods in Section 5.1. We then move on to the real-world dataset in Section 5.2 to show a method of learning the distribution of inward probabilities from online social networks, and the performance of partitioned sampling on the real-world graph based on the learned inward probabilities.

In our experiment, when the parameters of VIO model (i.e. weighted adjacency matrix A, people's inward probabilities $p_1, p_2, \ldots, p_n$, updating rates of people's opinions



(a) Synthetic graph (100 nodes) (b) Synthetic graph (10,000 nodes) (c) Synthetic graph (10,000 nodes)

Figure 2: Experimental results of synthetic graphs.



(a) Distribution of inward probabilities (b) Weibo graph with $\lambda_i = 1$ (c) Weibo graph with λ_i set to the number of v_i's tweets

Figure 3: Performance of various algorithms.

$\lambda_1, \lambda_2, \ldots, \lambda_n$, and the mean of innate opinion μ) are given, the experiment are done by (a) calculating the correlations of every pair of nodes by Theorem 1, (b) running the partitioning algorithms[2] to obtain the partition candidate, and (c) computing the expected variance $\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$[3] by Theorem 3. In both synthetic and real-world datasets, we set μ to 0.5. Notice that the value of μ has no effect on the results.

5.1 Synthetic Dataset

In our synthetic experiments, we use the hidden partition model [3] to generate the undirected graphs, which aims at resembling the community structure in real-world social networks. It is specified by four parameters: the number of vertices n, the number of hidden partitions k, the inter-partition and intra-partition edge probabilities p_H and p_L, respectively. First, we assign each node to one of the k hidden partitions uniformly at random. Next, we independently connect each pair of nodes in the same hidden partition with probability p_H, and two nodes in different partitions with probability $p_L < p_H$. We generate two different sizes of hidden partition graphs. The small one is used to compare the sampling quality among the following three different sampling methods: naive sampling (Naive), partitioned sampling using greedy partitioning (Greedy), and partitioned sampling using SDP partitioning (SDP). We use a small graph because SDP is infeasible to run in large graphs. We then move on to the large hidden partition graphs and only run Greedy and Naive for these graphs. We study the impact of inward probabilities and graph structure on the performance of Greedy on those graphs, respectively. For the synthetic graphs, we set opinion updating rates λ_i to 1 for all $i \in [n]$.

Small synthetic graph. The small hidden partition graphs we generate includes 100 nodes and 20 hidden partitions. Probability p_H and p_L are set to 0.9 and 0.01, respectively. The inward probability of each node is randomly chosen from $[0, 0.01]$. We range the sample size r from 2 to 14, and the expected sample variance is shown in Figure 2(a). We put sample size on y-axis to make it easier to see the savings on the sample size under the same expected sample variance.

When the sample size r is small (i.e. less than 6), the performance of SDP and Greedy are similar to each other, and both better than Naive. When the sample size r increases,

[2]To solve the SDP programming in SDP partitioning algorithm, we used CVX which is a package for specifying and solving convex programs [12, 13].

[3]Each randomized partitioning algorithm was run 10 times, and we took the average of the expected variance as the result.

the performance of Greedy becomes much better than Naive, and the performance of SDP starts getting worse but it is still better than Naive. Specifically, if we fix $\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$ to be 0.008, Greedy needs 9 samples, and SDP needs 11 samples, while Naive needs 13 samples. This suggests that by using our partitioned sampling method, we can save 31% of samples while achieving the same sampling quality.

Large synthetic graphs. For the large synthetic graph with 10k nodes and 500 hidden partitions, SDP is no longer feasible, thus we compare the improvement of Greedy against Naive. We ran Greedy and Naive using different sample sizes ($r = 250$ and $r = 500$), varying the inward probabilities and p_H/p_L, to observe the improvement of expected sample variance under different graph clustering and inward tendency settings.

In Figure 2(b), we set all nodes' inward probabilities to 0.05 and p_H to 1, and range p_L from 10^{-1} to 10^{-6}. The improvement on y-axis means the improvement of expected sample variance from Naive to Greedy. The result shows that the larger p_H/p_L (more apparent clustering) indicates the better performance of our partitioned sampling. When p_H/p_L increases from 10^2 to 10^5, the improvement of expected sample variance enhances rapidly. When p_H/p_L is too large (i.e. larger than 10^5), the improvement of expected sample variance becomes saturated. This is because the number of edges which cross different hidden partitions are very few so that it decreases rather slowly and the graph structure is almost unchanged when p_H/p_L increases further.

In Figure 2(c), we set p_H to 1 and p_L to be 10^{-5} to generate the hidden partition graph. For this graph, we set all nodes' inward probabilities to be identical, varying from 0.02 to 0.8. The result shows that the lower inward probability indicates the better performance of our partitioned sampling. When the inward probability is small, the improvement of expected sample variance increases rapidly. This is because a lower inward probability means people interact more with their friends and thus their opinions are correlated more significantly.

According to the above two experiments, we conclude that the larger p_H/p_L and the lower inward probability make people's opinions clustered and strongly correlated inside the clusters, and our partitioned sampling method works better for these cases.

5.2 Real-World Dataset

The real-network dataset we use is the online social network data from Sina Weibo[4] [20], 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). For this dataset, we first need to learn the distribution of people's inward probabilities.

5.2.1 Distribution of Inward Probabilities

In order to observe the evolution of opinions for a specific topic of interest, We manually choose 12 specific topics (e.g. Microsoft, iPhone, etc.), and extract all tweets from the Weibo dataset related to these topics (simply using keyword based classifier). We then run each tweet through a sentiment analyzer [18] to obtain binary opinion values (positive/negative). Thus we get a series of opinions for each user at discrete time corresponding to each topic. For each topic, we select those users who published opinions at least 4 times, and regard their first opinions as their innate opinions $f_1(0)$, $f_2(0)$, ..., $f_n(0)$ and treat the average of the rest opinions as their expected opinions $\mathbb{E}_M[f_1]$, $\mathbb{E}_M[f_2]$, ..., $\mathbb{E}_M[f_n]$ in the in steady state state. We then collect their relationships and form a subgraph for the corresponding topic.

Recall the definition of matrix Q (Definition 2), and it is easy to see that

$$\begin{pmatrix} \mathbb{E}_M[f_1] \\ \mathbb{E}_M[f_2] \\ \cdots \\ \mathbb{E}_M[f_n] \end{pmatrix} = Q \begin{pmatrix} f_1(0) \\ f_2(0) \\ \cdots \\ f_n(0) \end{pmatrix} \quad (\text{or } \mathbb{E}_M[\vec{f}] = Q\vec{f}(0)).$$

Thus we can estimate the inward probabilities by solving the following programming

$$\begin{aligned} &\text{Minimize} \quad \left\| E_M[\vec{f}] - Q\vec{f}(0) \right\|, \\ &\text{Subject to} \quad 0 \le p_i \le 1, \forall i \in [n], \end{aligned}$$

and we use gradient descent method to handle above programming.

We estimate the inward probabilities under the 12 topics respectively, and Figure 3(a) shows the distribution of inward probabilities for three topics, namely Spring Festival Gala (68 users), Microsoft (66 users) and iPhone (59 users), and the results for other topics are similar. The distribution for these three different topics are quite similar: (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 many people are affected by other people's opinions. We manually look up those users who locate in $[0.8, 1]$, and find that a large number of them are media accounts and verified users. This matches our intuition that those users alway take effort to spread their own opinions on the web but less likely to adopt other people's opinions, hence they should have large inward probabilities.

5.2.2 Performance of Partitioned Sampling

In this section, we show the performance of Greedy compared to Naive on the real-world graph. We randomly select 22,000 users from the Weibo dataset, and remove the users who do not follow anyone, iteratively. Then we get our large Weibo graph including 10,975 nodes and 25,236 directed edges. 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 from 1/1/2013 to 1/1/2014 in the Weibo dataset. The users' inward probabilities are set in the following way so that it follows the distribution we learned in the previous section. We sort all the inward probabilities learned in Section 5.2.1 among 12 topics, denoted as $\hat{p}_1 < \hat{p}_2 < \cdots < \hat{p}_k$. For each user in the large Weibo graph, we select an integer i from $\{1, 2, \ldots, k+1\}$ uniformly at random, and set her inward probability to a random real number in the following interval

$$\begin{cases} [0, \hat{p}_1], & \text{if } i = 1, \\ [\hat{p}_k, 1], & \text{if } i = k+1, \\ [\hat{p}_{i-1}, \hat{p}_i], & \text{others.} \end{cases}$$

Since there are some $\hat{p}_i$ values that are zeros, we will have users with zero inward probability. For these users, we use

[4] http://weibo.com

a very small value 10^{-10} in our simulation since our computation of the VIO model requires inward probability to be greater than zero.

Figure 3(b) and 3(c) show the experimental result on the Weibo graph with all $\lambda_i = 1$ and λ_i set to the number of v_i's tweets, respectively. The improvement of Greedy against Naive with two different updating rate settings are similar. In particular, if we fix $\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$ to be 2.1×10^{-4}, Greedy needs 920 samples while Naive needs 1074 samples (saving 14.3%) in Figure 3(b), and Greedy needs 923 samples while Naive needs 1074 samples (saving 14.1%) in Figure 3(c).

The results indicate that the sample size saving is more apparent when the expected sample variance is getting smaller (i.e. sampling quality requirement is higher). The figures also indicate that our partitioned sampling method is robustly better than naive sampling method regardless of the updating rate settings. The results are consistent with the results from synthetic graphs in demonstrating the better performance of partitioned sampling method.

In summary, our results on the real-world data show that real-world social networks do exhibit opinion correlations and clusterings that can enable more efficient sampling through the partitioned sampling method. Our results on the synthetic data further show that when graph clustering and social interaction are stronger, the benefit of partitioned sampling could be higher.

6. DISCUSSION

We now provide further discussions on several aspect of our partitioned sampling approach.

Correlation learning and balanced greedy algorithm. As discussed in Section 4, our partitioned sampling method (in particular Theorem 3) can be applied with any method of learning opinion correlations between pairs of users. Thus, besides using VIO, one can potentially apply other models to model and learn opinion correlations. Alternatively, if enough data is available, one may learning correlations directly from historical opinion data of all users.

If the learned correlations are not accurate, our partitioned sampling method using either SDP or greedy approach may lose its effectiveness. However, if we make a small modification of SDP partitioning algorithm to force the output partition to be a balanced partition, then by Lemma 4, the balanced greedy algorithm always achieves a better sampling quality than Naive, when the population size n is large enough, no matter whether our correlation estimation is accurate or not.

Objective function of the OPS problem. In the definition of OPS problem (Section 2), we use $\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$ as our objective function where M represents the randomness from the evolution model and S represents the randomness from sampling. Since our partitioning algorithms (i.e. Algorithm 4.1 and 4.2) are randomized algorithms, the randomness S can further be divided into two parts: the randomness from partitioning algorithms (denoted as P) and the randomness from sample selecting in each component (denoted as C). Thus the objective function of our OPS problem can be specifically written as $\mathbb{E}_M[\mathrm{Var}_{P,C}(\hat{f})]$. However, in our experimental evaluation (Section 5), we compute $\mathbb{E}_M[\mathrm{Var}_C(\hat{f})]$ for each partition and take average of them to be the "$\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$", which is strictly to be $\mathbb{E}_P\,\mathbb{E}_M[\mathrm{Var}_C(\hat{f})]$. In fact, they are the same as we show below. Notice that the randomness M and P are independent, thus $\mathbb{E}_M$ and $\mathbb{E}_P$ are commutative. Thus

$$
\begin{aligned}
&\mathbb{E}_P\,\mathbb{E}_M[\mathrm{Var}_C(\hat{f})] - \mathbb{E}_M[\mathrm{Var}_S(\hat{f})] \\
&\quad= \mathbb{E}_P\,\mathbb{E}_M\left[\mathbb{E}_C[\hat{f}^2] - \mathbb{E}_C[\hat{f}]^2\right] \\
&\qquad - \mathbb{E}_M\left[\mathbb{E}_P\,\mathbb{E}_C[\hat{f}^2] - \left[\mathbb{E}_P\,\mathbb{E}_C[\hat{f}]\right]^2\right] \\
&\quad= \mathbb{E}_M\,\mathbb{E}_P\,\mathbb{E}_C[\hat{f}^2] - \mathbb{E}_M\,\mathbb{E}_P\left[\mathbb{E}_C[\hat{f}]\right]^2 \\
&\qquad - \mathbb{E}_M\,\mathbb{E}_P\,\mathbb{E}_C[\hat{f}^2] + \mathbb{E}_M\left[\mathbb{E}_P\,\mathbb{E}_C[\hat{f}]\right]^2 \\
&\quad= \mathbb{E}_M\left[\left(\mathbb{E}_P\,\mathbb{E}_C[\hat{f}]\right)^2 - \mathbb{E}_P\left[\mathbb{E}_C[\hat{f}]\right]^2\right] \\
&\quad= \mathbb{E}_M\,\mathrm{Var}_P\left(\mathbb{E}_C[\hat{f}]\right) \\
&\quad= \mathbb{E}_M\,\mathrm{Var}_P(\bar{f}) \qquad (3) \\
&\quad= 0.
\end{aligned}
$$

Equation (3) holds because $\hat{f}$ is an unbiased estimate for any partition $\mathcal{P}$. Thus $\mathbb{E}_P\,\mathbb{E}_M[\mathrm{Var}_C(\hat{f})]$ and $\mathbb{E}_M[\mathrm{Var}_S(\hat{f})]$ are the same, and we do not distinguish them in the paper.

7. CONCLUSION AND FUTURE WORK

In this paper, we propose the partitioned sampling method based on the VIO model to achieve the better sampling quality, by proposing the SDP and greedy partitioning algorithms. We also apply the partitioned sampling method on both synthetic and real-world graphs, and show that partitioned sampling achieves effective improvement on sample quality (or equivalently savings on sample size).

There are a number of open problems and future directions one may pursue. For example, one may further enrich the VIO model to allow (a) non-identical innate opinion distributions if partial knowledge about individuals' innate opinion tendency is available, or (b) negative relationships as modeled in [14] so as to include negative correlations, and study the OPS problem under these models. Another direction is to improve learning the parameters of VIO model, or in general to extract opinion correlations in social networks from real-world data. Finally, the current approach is based on learning all pairs of node correlations, which is infeasible for very large graphs with tens or hundreds of millions of nodes. Thus how to bypass all-pair-correlation computation and achieve highly scalable partitioned sampling algorithm is an important task we plan to pursue in the future.

APPENDIX

A. MATHEMATICAL PROOFS

PROPOSITION 1. *Partitioned sampling given in Algorithm 2.1 is unbiased.*

PROOF. For any partition $\mathcal{P} = \{(V_1, r_1), \ldots, (V_K, r_K)\}$,

$$\begin{aligned}\mathbb{E}_S\left[\hat{f}_{part}(\mathcal{P})\right] &= \frac{1}{|V|}\sum_{k=1}^{K}|V_k|\,\mathbb{E}_S\left[\hat{f}_{naive}(V_k, r_k)\right] \\ &= \frac{1}{|V|}\sum_{k=1}^{K}|V_k|\sum_{v_i\in V_k}\frac{f_i}{|V_k|} \\ &= \frac{\sum_{i=1}^{n} f_i}{|V|} \\ &= \bar{f}.\end{aligned}$$

Notice that naive sampling in V_k with no replacement is unbiased, thus $\mathbb{E}_S[\hat{f}_{naive}(V_k, r_k)]$ is equal to the mean of f_i in V_k. Thus partitioned sampling is unbiased. □

LEMMA 1. *When $p_i > 0$ for all $i \in [n]$, the VIO model has a unique joint distribution for the final expressed opinions in steady state.*

PROOF. The opinion evolution can be viewed as a Markov chain. Each possible assignment of $v_1, v_2, \ldots, v_n$'s expressed opinions forms one state and the initial state of the Markov chain is $(f_1(0), f_2(0), \ldots, f_n(0))$. At each Poisson arrival time, the transition from one state to another represents the change of the opinion assignment. Thus the state space consists of all the states reachable from the initial state. The VIO model has a unique steady state distribution for the final expressed opinions if and only if the Markov chain has a unique stationary distribution. In order to prove the existence of the unique stationary distribution of the Markov chain, we only need to prove that the Markov chain is irreducible and aperiodic [17]. Notice that each state in the state space can be reached from the initial state. Meanwhile, each state in the state space can return to the initial state by each node updating its expressed opinion to the innate opinion which happens with a positive probability. This means that any two states in the state space are connected, indicating the irreducibility of the Markov chain. In addition, the initial state is aperiodic since it has a self-loop in the state transition graph (with probability at least $\sum_i^n p_i/n > 0$). An irreducible Markov chain is aperiodic if there exists one aperiodic state. Therefore, the Markov chain is irreducible and aperiodic, with the unique stationary distribution being reached after long enough time. □

LEMMA 2. *The expected expressed opinion of each node in steady state is equal to the mean of innate opinion, that is $\mathbb{E}_M[f_i] = \mu$ for all $i \in [n]$.*

PROOF. We prove this lemma by proving a stronger statement: given any $t \geq 0$, $\mathbb{E}_M[f_i(t)] = \mu$ for all $i \in [n]$. Namely, we want to prove that at any time t, each node's expected expressed opinion is equal to the mean of innate opinion.

The proof is by induction on time t.

In the initial state, each node's expressed opinion(also innate opinion) is generated from an i.i.d. distribution and the above statement holds.

Now suppose the statement holds before time t. It still holds before the next Poisson arrival among all nodes. Suppose the next Poisson arrival comes at time τ and its corresponding updating node is v_i. At this Poisson arrival time $\tau > t$, v_i updates its expressed opinion based on its innate opinion and one of its neighbors' expressed opinions. Notice that the expectations of both its innate opinion $f_i(0)$ and all its neighbors' expressed opinions $f_j(\tau)$ are equal to μ by the inductive assumption, namely, $\mathbb{E}_M[f_i(0)] = \mu$ and $\mathbb{E}_M[f_j(\tau)] = \mu$ for all $v_j \in N_i$ where N_i is the set of v_i's neighbors. Thus the expectation of v_i's updated expressed opinion $\mathbb{E}_M[f_i(\tau)]$ is still equal to μ. Moreover, other nodes' expected expressed opinions remain equal to μ upon time τ.

Thus by induction, at any time t, each node's expected expressed opinion is always equal to μ. □

LEMMA 3. *$\mathbb{P}\left[\mathcal{I}_{ij}^l\right]$, $i, j, l \in [n]$ is the unique solution of the following linear equation system:*

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

Q is computed by

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

where $P = \mathrm{diag}(p_1, p_2, \ldots, p_n)$ and $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ are two diagonal matrices, and matrix $I - (I-P)\,D^{-1}A$ is invertible when $p_i > 0$ for all $i \in [n]$.

PROOF. (a) Recall from Definition 2 that $\mathcal{I}_{ij}^l$ denotes the event that two random walks starting from v_i and v_j at time $t = \infty$ eventually meet and the first node they meet at is $v_l \in V$. This event consists of two steps: 1) The walker at v_i (or v_j) moves to one of its neighbor v_a (or v_b); 2) two random walks starting from v_a (or v_b) and v_j (or v_i) eventually meet and the first node they meet at is v_l. The probability that the walker at v_i (resp. v_j) make a movement is proportional to v_i's (resp. v_j's) Poisson rate, that is $\lambda_i/(\lambda_i+\lambda_j)$ (resp. $\lambda_j/(\lambda_i+\lambda_j)$). Thus when $i \neq j$, $\mathbb{P}\left[\mathcal{I}_{ij}^l\right]$ can be calculated by the following recursion:

$$\begin{aligned}\mathbb{P}\left[\mathcal{I}_{ij}^l\right] = &\sum_{a=1}^{n}\frac{\lambda_i}{\lambda_i+\lambda_j}\frac{(1-p_i)A_{ia}}{d_i}\mathbb{P}\left[\mathcal{I}_{aj}^l\right] \\ &+ \sum_{b=1}^{n}\frac{\lambda_j}{\lambda_i+\lambda_j}\frac{(1-p_j)A_{jb}}{d_j}\mathbb{P}\left[\mathcal{I}_{ib}^l\right].\end{aligned}$$

When $i = j$, $\mathbb{P}\left[\mathcal{I}_{ij}^l\right]$ can be determined by the following boundary conditions:

$$\mathbb{P}\left[\mathcal{I}_{ij}^l\right] = \begin{cases} 0, & i = j \neq l, \\ 1, & i = j = l. \end{cases}$$

By combining the recursive equations and the boundary conditions, we have the following linear equations:

$$\mathbb{P}\left[\mathcal{I}_{ij}^l\right] = \begin{cases} 0, & i = j \neq l, \\ 1, & i = j = l, \\ \sum_{a=1}^{n} \frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i}\mathbb{P}[\mathcal{I}_{aj}^l] & \\ \quad + \sum_{b=1}^{n} \frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j}\mathbb{P}[\mathcal{I}_{ib}^l], & i \neq j. \end{cases} \tag{4}$$

(The above proof follows the idea in [19].)

Now we show that the linear equations has a unique solution.

For a fixed l, the equations for all terms $\mathbb{P}\left[\mathcal{I}_{ij}^{l}\right]$ such that $i \neq j$ form a linear sub-system of $\binom{n}{2}$ variables and $\binom{n}{2}$ equations. Therefore, we can solve the whole linear system (4) by solving n separated linear sub-systems. Each linear sub-system corresponds to a value of l, and it can be solved in $O\left(\binom{n}{2}^3\right) = O\left(n^6\right)$[5] , thus the original linear system (4) can be solved in time $n \cdot O\left(n^6\right) = O\left(n^7\right)$, as mentioned in Section 3.2.

Now, we show that there exists the unique solution for each linear sub-system.

Each equation in the linear sub-system can be written as

$$\begin{aligned}\mathbb{P}\left[\mathcal{I}_{ij}^{l}\right] = &\sum_{a\neq j}\frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i}\mathbb{P}\left[\mathcal{I}_{aj}^{l}\right]\\ &+\sum_{b\neq i}\frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j}\mathbb{P}\left[\mathcal{I}_{ib}^{l}\right]\\ &+\frac{\lambda_i(1-p_i)A_{ij}}{(\lambda_i+\lambda_j)d_i}\cdot 1_{j=l}+\frac{\lambda_j(1-p_j)A_{ji}}{(\lambda_i+\lambda_j)d_j}\cdot 1_{i=l}.\end{aligned}$$

Let $k = h(i,j) = (i-1)n + j$, then we have a bijection h of subscript between integer k and ordered pair (i,j) where $i < j$. We can write these equations in matrix form:

$$I\vec{x} = M_1\vec{x} + M_2\vec{x} + \vec{b}$$

where $\vec{x}$ is the $\binom{n}{2}\times 1$ vector whose k-th element is $\mathbb{P}\left[\mathcal{I}_{ij}^{l}\right]$; M_1 is the $\binom{n}{2}\times\binom{n}{2}$ matrix whose $(k, h(a,j))$ entry is $\frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i}$; M_2 is the $\binom{n}{2}\times\binom{n}{2}$ matrix whose $(k,h(i,b))$ entry is $\frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j}$; $\vec{b}$ is the $\binom{n}{2}\times 1$ vector whose k-th element is $\frac{\lambda_i(1-p_i)A_{ij}}{(\lambda_i+\lambda_j)d_i}\cdot 1_{j=l} + \frac{\lambda_j(1-p_j)A_{ji}}{(\lambda_i+\lambda_j)d_j}\cdot 1_{i=l}$.

If $I - M_1 - M_2$ is non-singular, each linear sub-system has a unique solution $(I-M_1-M_2)^{-1}\vec{b}$. In fact, for any row s of $I-M_1-M_2$, let $(i,j) = h^{-1}(s)$, and

$$\begin{aligned}\sum_{t\neq s}|I-M_1-M_2|_{st} &= \sum_{a\neq j}\frac{\lambda_i(1-p_i)A_{ia}}{(\lambda_i+\lambda_j)d_i}+\sum_{b\neq i}\frac{\lambda_j(1-p_j)A_{jb}}{(\lambda_i+\lambda_j)d_j}\\ &= \frac{\lambda_i(1-p_i)}{(\lambda_i+\lambda_j)}\sum_{a\neq j}\frac{A_{ia}}{d_i}+\frac{\lambda_j(1-p_j)}{(\lambda_i+\lambda_j)}\sum_{b\neq i}\frac{A_{jb}}{d_j}\\ &\leq \frac{\lambda_i(1-p_i)}{(\lambda_i+\lambda_j)}+\frac{\lambda_j(1-p_j)}{(\lambda_i+\lambda_j)}\\ &< \frac{\lambda_i}{(\lambda_i+\lambda_j)}+\frac{\lambda_j}{(\lambda_i+\lambda_j)}\\ &= |I-M_1-M_2|_{ss}.\end{aligned}$$

Thus, $I - M_1 - M_2$ is strictly diagonally dominant, and it is non-singular [1]. Since each linear sub-system has one unique solution, the whole linear system (4) also does.

(b) The probability of a walker from v_i walking to v_j is $p_{ij}^{(1)} = (1-p_i)A_{ij}/d_i$. Note here the statement is conditioned on the current Poisson arrival is v_i. So we have a matrix form $P_{VV} = (I-P)D^{-1}A$ whose (i,j) is $p_{ij}^{(1)}$. Therefore, the probability of walking from v_i to v_j in exactly l steps is the (i,j) entry of $(P_{VV})^l$.

By definition of our model, the probability of v_j walking to v_j' (being absorbed) is p_j. Thus the matrix Q whose (i,j) entry is the probability of transition from v_i to v_j' can be calculated by

$$Q = \sum_{l=0}^{\infty}(P_{VV})^l P = (I-P_{VV})^{-1}P = \left(I-(I-P)D^{-1}A\right)^{-1}P.$$

Now we show that $I - P_{VV}$ is invertible. The (i,j) entry of $I - P_{VV}$ is

$$\begin{cases}1, & \text{if } i=j,\\ -\frac{1-p_i}{d_i}A_{ij}, & \text{if } i\neq j.\end{cases}$$

For any row i of $I - P_{VV}$, the sum of absolute values of its non diagonal elements can be written as $\sum_{j\neq i}\frac{1-p_i}{d_i}A_{ij} = (1-p_i)(1-\frac{A_{ii}}{d_i})$, and it is strictly less than the absolute value of i-th diagonal elements $|I-P_{VV}|_{ii} = 1$. Thus $I - P_{VV}$ is strictly diagonally dominant, and it is non-singular [1]. □

THEOREM 1. *For any $i,j\in[n]$, correlation $\mathrm{Cor}_M(f_i,f_j)$ 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, $\mathrm{Cor}_M(f_i,f_j)$ can be computed by*

$$\mathrm{Cor}_M(f_i,f_j) = \sum_{l=1}^{n}\mathbb{P}\left[\mathcal{I}_{ij}^{l}\right]\left(1-\sum_{k=1}^{n}Q_{lk}^2\right)+\sum_{k=1}^{n}Q_{ik}Q_{jk},$$

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

PROOF. (a) In this part, we show that $\mathrm{Cor}_M(f_i,f_j)$ 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'. In the proof, we split the randomness M into two parts: We use O to denote the randomness of innate opinions which are generated by an i.i.d. distribution, and we use E to denote the randomness from the opinion evolution.

When $i=j$, obviously we have $\mathrm{Cor}_M(f_i,f_j)=1$. In this case, the two random walks' paths coincide, thus they are absorbed by the same node in V' with probability 1.

When $i\neq j$, according to the definition of correlation,

$$\begin{aligned}\mathrm{Cor}_M(f_i,f_j) &= \frac{\mathbb{E}_M[f_if_j]-\mathbb{E}_M[f_i]\,\mathbb{E}_M[f_j]}{\sqrt{\mathrm{Var}_M[f_i]\,\mathrm{Var}_M[f_j]}}\\ &= \frac{\mathbb{E}_M[f_if_j]-\mu^2}{\mu-\mu^2}.\end{aligned}$$

The second equality holds because for any $i\in[n]$,

$$\mathrm{Var}_M[f_i] = \mathbb{E}_M[f_i^2]-\mathbb{E}_M[f_i]^2 = \mathbb{E}_M[f_i]-\mathbb{E}_M[f_i]^2 = \mu-\mu^2.$$

Next, we need to calculate $\mathbb{E}_M[f_if_j]$, which is the probability that two random walkers starting from v_i and v_j walk to the nodes in V' whose original opinions are 1. This event consists of two cases: Two random walkers move to the same absorbing node, or two distinct absorbing nodes. Thus we can calculate $\mathbb{E}_M[f_if_j]$ by adding them together.

Let $\mathcal{M}_{i,j}^{p,q}$ be the event that in the coalescing random walk on $\overline{G}$, a random walker starting from v_i is absorbed by v_p', while another random walker starting from v_j is absorbed by v_q'. Note that $\mathcal{M}_{i,j}^{p,q}$ is measurable under randomness E. It only depends on the structure of $\overline{G}$ and is independent of the initial value in V':

$$\mathbb{P}_E\left[\mathcal{M}_{i,j}^{p,q}\mid\vec{f}(0)\right] = \mathbb{P}_E\left[\mathcal{M}_{i,j}^{p,q}\right].$$

[5] n-variable linear system can be solved in time $O\left(n^3\right)$.

Thus $\mathbb{E}_M[f_i f_j]$ can be written as:

$$\begin{aligned}
&\mathbb{E}_M[f_i f_j] \\
&\quad = \mathbb{E}_{O,E}[f_i f_j] \\
&\quad = \sum_{p\neq q} \mathbb{P}_{O,E}\left[\mathcal{M}_{i,j}^{p,q} \mid f_p(0)f_q(0)=1\right]\mathbb{P}_{O,E}\left[f_p(0)f_q(0)=1\right] \\
&\qquad\quad + \sum_{p=1}^{n} \mathbb{P}_{O,E}\left[\mathcal{M}_{i,j}^{p,p} \mid f_p(0)=1\right]\mathbb{P}_{O,E}\left[f_p(0)=1\right] \\
&\quad = \sum_{p\neq q} \mathbb{P}_{E}\left[\mathcal{M}_{i,j}^{p,q}\right]\mathbb{P}_{O}\left[f_p(0)f_q(0)=1\right] \\
&\qquad\quad + \sum_{p=1}^{n} \mathbb{P}_{E}\left[\mathcal{M}_{i,j}^{p,p}\right]\mathbb{P}_{O}\left[f_p(0)=1\right] \\
&\quad = \mu^2 \sum_{p\neq q} \mathbb{P}_{E}\left[\mathcal{M}_{i,j}^{p,q}\right] + \mu \sum_{p=1}^{n} \mathbb{P}_{E}\left[\mathcal{M}_{i,j}^{p,p}\right] \\
&\quad = \mu \sum_{p=1}^{n} \mathbb{P}_{E}\left[\mathcal{M}_{i,j}^{p,p}\right] + \mu^2\left(1 - \sum_{p=1}^{n} \mathbb{P}_{E}\left[\mathcal{M}_{i,j}^{p,p}\right]\right) \\
&\quad =: \mu p_{same}(i,j) + \mu^2(1-p_{same}(i,j)).
\end{aligned}$$

In the last equation, we use $p_{same}(i,j)$ to denote the probability that two coalescing random walks starting from v_i and v_j end at the same node in V'. Thus

$$\begin{aligned}
\mathrm{Cor}_M(f_i,f_j) &= \frac{\mathbb{E}_M[f_i f_j]-\mu^2}{\mu-\mu^2} \\
&= \frac{\left[\mu p_{same}(i,j)+\mu^2(1-p_{same}(i,j))\right]-\mu^2}{\mu-\mu^2} \\
&= p_{same}(i,j).
\end{aligned}$$

This means that $\mathrm{Cor}_M(f_i,f_j)$ is equal to the probability that two coalescing random walks starting from v_i and v_j end at the same absorbing node in V'.

(b) We now calculate p_{same} in this part. Let $\mathcal{H}_{ij}^k$ be the event that two coalescing random walks starting from v_i and v_j are both absorbed by node v'_k without meeting each other at a node in V. According to the definitions of events $\mathcal{M}_{i,j}^{p,q}$, $\mathcal{I}_{ij}^l$ and $\mathcal{H}_{ij}^k$, we have

$$\mathbb{P}_E\left[\mathcal{M}_{i,j}^{k,k}\right] = \sum_{l=1}^{n} \mathbb{P}\left[\mathcal{I}_{ij}^l\right] Q_{lk} + \mathbb{P}\left[\mathcal{H}_{ij}^k\right] \tag{5}$$

where Q_{lk} is the probability that a random walker starting from node v_l at time ends at $v'_k \in V'$.

Notice that $Q_{ik}Q_{jk}$ represents the probability that two *non*-coalescing random walks starting from v_i and v_j end at node v'_k, thus it can be written as:

$$Q_{ik}Q_{jk} = \sum_{l=1}^{n} \mathbb{P}\left[\mathcal{I}_{ij}^l\right] Q_{lk}^2 + \mathbb{P}\left[\mathcal{H}_{ij}^k\right]. \tag{6}$$

Combining Equation (5) and (6),

$$\mathbb{P}_E\left[\mathcal{M}_{i,j}^{k,k}\right] = \sum_{l=1}^{n} \mathbb{P}\left[\mathcal{I}_{ij}^l\right](Q_{lk}-Q_{lk}^2) + Q_{ik}Q_{jk}.$$

Therefore,

$$\begin{aligned}
\mathrm{Cor}_M(f_i,f_j) &= \sum_{k=1}^{n} \mathbb{P}_E\left[\mathcal{M}_{i,j}^{k,k}\right] \\
&= \sum_{k=1}^{n}\left(\sum_{l=1}^{n} \mathbb{P}\left[\mathcal{I}_{ij}^l\right](Q_{lk}-Q_{lk}^2) + Q_{ik}Q_{jk}\right) \\
&= \sum_{l=1}^{n} \mathbb{P}\left[\mathcal{I}_{ij}^l\right]\left(1-\sum_{k=1}^{n} Q_{lk}^2\right) + \sum_{k=1}^{n} Q_{ik}Q_{jk}.
\end{aligned}$$

This finishes the proof. □

LEMMA 4. *Given a graph with $n \geq 2$ vertices, partitioned sampling using any balanced complete partition $\mathcal{P}$ of the graph, is better than naive sampling from the graph (after ignoring an $o(1)$ term). Specifically,*

$$\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P})) < \mathrm{Var}_S(\hat{f}_{naive}) + \frac{3}{2n}.$$

PROOF. For the social graph to be sampled, let $\bar{f}$ be the average opinion of the whole graph $\bar{f} = \sum_{i=1}^{n} f_i/n$. The estimate of naive sampling is $\hat{f}_{naive} = \sum_{k=1}^{r} f_{s_k}/r$ where r is the sample size and v_{s_k} ($k \in [r]$) are the sampled nodes, and the estimate of complete partitioned sampling is denoted by $\hat{f}_{part}(\mathcal{P}) = \sum_{k=1}^{r} n_k f_{s_k}/n$ where n_k is the size of component V_k and $\sum_{k=1}^{r} n_k = n$.

The variance of naive sampling without replacement is:

$$\begin{aligned}
\mathrm{Var}_S\left(\hat{f}_{naive}\right) &= \mathbb{E}_S[\hat{f}_{naive}^2] - \mathbb{E}_S[\hat{f}_{naive}]^2 \\
&= \frac{1}{r^2}\mathbb{E}_S\left[\left(\sum_{k=1}^{r} f_{s_k}\right)^2\right] - \bar{f}^2 \\
&= \frac{1}{r^2}\sum_{k=1}^{r}\mathbb{E}_S\left[f_{s_k}^2\right] + \frac{1}{r^2}\sum_{k\neq l}\mathbb{E}_S[f_{s_k}f_{s_l}] - \bar{f}^2.
\end{aligned}$$

Notice that $f_{s_k} \in \{0,1\}$,

$$\mathbb{E}_S\left[f_{s_k}^2\right] = \mathbb{E}_S\left[f_{s_k}\right] = \bar{f},$$

and

$$\mathbb{E}_S[f_{s_k}f_{s_l}] = \mathbb{P}_S[f_{s_k}=1]\mathbb{P}_S[f_{s_l}=1|f_{s_k}=1] = \bar{f}\cdot\frac{n\bar{f}-1}{n-1}.$$

Thus

$$\begin{aligned}
\mathrm{Var}_S\left(\hat{f}_{naive}\right) &= \frac{1}{r^2}\sum_{k=1}^{r}\mathbb{E}_S\left[f_{s_k}^2\right] + \frac{1}{r^2}\sum_{k\neq l}\mathbb{E}_S[f_{s_k}f_{s_l}] - \bar{f}^2 \\
&= \frac{1}{r^2}r\bar{f} + \frac{1}{r^2}r(r-1)\bar{f}\cdot\frac{n\bar{f}-1}{n-1} - \bar{f}^2 \\
&= \frac{\bar{f}(1-\bar{f})}{r}\frac{n-r}{n-1}.
\end{aligned} \tag{7}$$

The variance of partitioned sampling using complete partition $\mathcal{P}$ can be written as

$$\begin{aligned}
&\mathrm{Var}_S\left(\hat{f}_{part}(\mathcal{P})\right) \\
&\quad = \mathbb{E}_S\left[\hat{f}_{part}(\mathcal{P})^2\right] - \mathbb{E}_S\left[\hat{f}_{part}(\mathcal{P})\right]^2 \\
&\quad = \mathbb{E}_S\left[\sum_{k=1}^{r}\frac{n_k^2}{n^2}f_{s_k}^2 + \sum_{k\neq l}\frac{n_k n_l}{n^2}f_{s_k}f_{s_l}\right] \\
&\qquad - \left(\sum_{k=1}^{r}\frac{n_k}{n}\mathbb{E}_S\left[f_{s_k}\right]\right)^2 \\
&\quad = \sum_{k=1}^{r}\frac{n_k^2}{n^2}\left(\mathbb{E}_S\left[f_{s_k}^2\right] - \mathbb{E}_S\left[f_{s_k}\right]^2\right)
\end{aligned}$$

$$= \sum_{k=1}^{r} \left(\frac{n_k}{n}\right)^2 \mathrm{Var}_S\left(f_{s_k}\right)$$
$$= \sum_{k=1}^{r} \bar{f}_k\left(1-\bar{f}_k\right)\left(\frac{n_k}{n}\right)^2.$$

where $\bar{f}_k$ is the average opinion of the k-th component. We define δ_k to be $\frac{n_k}{n} - \frac{1}{r}$ such that $\sum_{k=1}^{r}\delta_k = 0$. Thus

$$\begin{aligned}
&\mathrm{Var}_S\left(\hat{f}_{part}(\mathcal{P})\right) - \mathrm{Var}_S\left(\hat{f}_{naive}\right) \\
&= \sum_{k=1}^{r} \bar{f}_k(1-\bar{f}_k)\left(\frac{n_k}{n}\right)^2 - \frac{n-r}{r(n-1)}\bar{f}\left(1-\bar{f}\right) \\
&= \sum_{k=1}^{r} \bar{f}_k\left(\frac{n_k}{n}\right)^2 - \sum_{k=1}^{r}\left(\frac{n_k\bar{f}_k}{n}\right)^2 \\
&\quad - \frac{n-r}{r(n-1)}\left[\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n} - \left(\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\right)^2\right] \\
&= \sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\left(\frac{1}{r}+\delta_k\right) - \frac{n-r}{r(n-1)}\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n} \\
&\quad - \sum_{k=1}^{r}\left(\frac{n_k\bar{f}_k}{n}\right)^2 + \frac{n-r}{r(n-1)}\left(\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\right)^2 \\
&= \sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\left(\frac{1-\frac{1}{r}}{n-1}+\delta_k\right) - \sum_{k=1}^{r}\left(\frac{n_k\bar{f}_k}{n}\right)^2 \\
&\quad + \frac{1}{r}\left(\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\right)^2 - \frac{1-\frac{1}{r}}{n-1}\left(\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\right)^2 \\
&= \frac{1-\frac{1}{r}}{n-1}\left(\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\right)\left(1-\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\right) \\
&\quad + \sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\delta_k + \frac{1}{r}\left(\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\right)^2 - \sum_{k=1}^{r}\left(\frac{n_k\bar{f}_k}{n}\right)^2 \\
&= \frac{1-\frac{1}{r}}{n-1}\bar{f}(1-\bar{f}) + \sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\delta_k \\
&\quad + \left[\frac{1}{r}\left(\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\right)^2 - \sum_{k=1}^{r}\left(\frac{n_k\bar{f}_k}{n}\right)^2\right].
\end{aligned}$$

According to Cauchy-Schwarz inequality,

$$\frac{1}{r}\left(\sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\right)^2 \le \sum_{k=1}^{r}\left(\frac{n_k\bar{f}_k}{n}\right)^2.$$

Thus we have

$$\mathrm{Var}_S\left(\hat{f}_{part}(\mathcal{P})\right) - \mathrm{Var}_S\left(\hat{f}_{naive}\right) \le \frac{1-\frac{1}{r}}{n-1}\bar{f}(1-\bar{f}) + \sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\delta_k.$$

Notice that for any balanced complete partition $\mathcal{P}$, we have $0 \le \delta_k < \frac{1}{n}$. Thus

$$\begin{aligned}
&\mathrm{Var}_S\left(\hat{f}_{part}(\mathcal{P})\right) - \mathrm{Var}_S\left(\hat{f}_{naive}\right) \\
&\le \frac{1-\frac{1}{r}}{n-1}\bar{f}(1-\bar{f}) + \sum_{k=1}^{r}\frac{n_k\bar{f}_k}{n}\delta_k \\
&< \frac{1}{n-1}\bar{f}(1-\bar{f}) + \bar{f}\cdot\sup\delta_k \\
&\le \frac{1}{(n-1)}\cdot\frac{1}{4} + \frac{1}{n} \\
&\le \frac{3}{2n}.
\end{aligned}$$

Thus we finish the proof. □

THEOREM 2. *Given a graph G and the VIO model on G, for any partition $\mathcal{P}$, partitioned sampling using the refined complete partition $\mathcal{P}'$ of $\mathcal{P}$ is better than partitioned sampling using the original partition $\mathcal{P}$ (after ignoring an $o(1)$ term). Specifically,*

$$\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}'))] < \mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))] + \frac{3}{2n}.$$

PROOF. From Lemma 4, if we take the expectation of randomness M, we have

$$\mathbb{E}_M\left[\mathrm{Var}_S\left(\hat{f}_{part}(\mathcal{P})\right)\right] < \mathbb{E}_M\left[\mathrm{Var}_S\left(\hat{f}_{naive}\right)\right] + \frac{3}{2n}.$$

For any partition strategy $\mathcal{P}$ with m components, we can find the balanced complete partition $\mathcal{P}_k^*$ for each component V_k such that

$$\mathbb{E}_M\left[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}_k^*))\right] < \mathbb{E}_M\left[\mathrm{Var}_S(\hat{f}_{naive}(V_k, r_k))\right] + \frac{3}{2n_k}.$$

Thus the refined complete partition $\mathcal{P}'$ of $\mathcal{P}$ satisfies that

$$\begin{aligned}
&\mathbb{E}_M\left[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}'))\right] \\
&= \sum_{k=1}^{m}\left(\frac{n_k}{n}\right)^2\mathbb{E}_M\left[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}_k^*))\right] \\
&< \sum_{k=1}^{m}\left(\frac{n_k}{n}\right)^2\mathbb{E}_M\left[\mathrm{Var}_S(\hat{f}_{naive}(V_k, r_k))\right] + \sum_{k=1}^{m}\frac{n_k^2}{n^2}\frac{3}{2n_k} \\
&= \mathbb{E}_M\left[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))\right] + \frac{3}{2n}.
\end{aligned}$$

This finishes the proof. □

THEOREM 3. *For any complete partition $\mathcal{P}$,*

$$\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))] = \frac{\mu(1-\mu)}{n^2}g_r(\mathcal{P}),$$

where μ is the mean of innate opinion. Thus, the best complete partition minimizes the cost function.

PROOF. We use s_k to represent the sample point selected in the k-th component V_k of partition $\mathcal{P}$ by complete partitioned sampling.

We define $\mathrm{Cut}(V_k, V_l) := \sum_{v_i\in V_k}\sum_{v_j\in V_l}\mathrm{Cor}_M(f_i, f_j)$ and $\mathrm{Cor}(V_k) := \sum_{i<j:v_i,v_j\in V_k}\mathrm{Cor}_M(f_i, f_j)$.

The estimate of complete partitioned sampling using partition $\mathcal{P}$ can be written as:

$$\hat{f}_{part}(\mathcal{P}) = \frac{\sum_{k=1}^{r} n_k f_{s_k}}{n},$$

where n_k is the size of V_k. Thus the sample variance of $\hat{f}_{part}(\mathcal{P})$ is:

$$\begin{aligned}
\mathrm{Var}_S\left[\hat{f}_{part}(\mathcal{P})\right] &= \mathbb{E}_S\left[\hat{f}_{part}^2(\mathcal{P})\right] - \mathbb{E}_S\left[\hat{f}_{part}(\mathcal{P})\right]^2 \\
&= \mathbb{E}_S\left[\left(\frac{\sum_{k=1}^{r} n_k f_{s_k}}{n}\right)^2\right] - \bar{f}^2 \\
&= \mathbb{E}_S\left[\frac{1}{n^2}\sum_{k=1}^{r} n_k^2 f_{s_k}^2 + \frac{2}{n^2}\sum_{k<l} n_k n_l f_{s_k} f_{s_l}\right] - \bar{f}^2
\end{aligned}$$

$$= \frac{1}{n^2}\sum_{k=1}^{r} n_k^2 \bar{f}_k + \frac{2}{n^2}\sum_{k<l} n_k n_l \bar{f}_k \bar{f}_l - \bar{f}^2.$$

where $\bar{f}$ is the average opinion of the entire population and $\bar{f}_k$ is the average opinion of the k-th component.

Then we take the expectation of each above item under the randomness M. Notice that $\mathbb{E}_M[f_i] = \mu$ for all $i \in [n]$. Thus we have

$$\mathbb{E}_M\left[\bar{f}_k\right] = \mathbb{E}_M\left[\frac{\sum_{v_i \in V_k} f_i}{n_k}\right] = \frac{\sum_{v_i \in V_k} \mu}{n_k} = \mu,$$

and

$$\begin{aligned}
\mathbb{E}_M\left[n_k n_l \bar{f}_k \bar{f}_l\right] &= \mathbb{E}_M\left[\left(\sum_{v_i \in V_k} f_i\right)\left(\sum_{v_j \in V_l} f_j\right)\right] \\
&= \sum_{v_i \in V_k}\sum_{v_j \in V_l} \mathbb{E}_M\left[f_i f_j\right] \\
&= \sum_{v_i \in V_k}\sum_{v_j \in V_l} \left[\mu^2 + \mu(1-\mu)\,\mathrm{Cor}_M(f_i, f_j)\right] \qquad (8) \\
&= n_k n_l \mu^2 + \mu(1-\mu)\,\mathrm{Cut}(V_k, V_l),
\end{aligned}$$

where Equation (8) holds because the definition of correlation

$$\mathrm{Cor}_M(f_i, f_j) = \frac{\mathbb{E}_M[f_i f_j] - \mathbb{E}_M[f_i]\,\mathbb{E}_M[f_j]}{\sqrt{\mathrm{Var}_M[f_i]\,\mathrm{Var}_M[f_j]}} = \frac{\mathbb{E}_M[f_i f_j] - \mu^2}{\mu - \mu^2}.$$

We also have

$$\begin{aligned}
\mathbb{E}_M\left[\bar{f}^2\right] &= \frac{1}{n^2}\,\mathbb{E}_M\left[\sum_{i=1}^{n}\sum_{j=1}^{n} f_i f_j\right] \\
&= \mu^2 + \frac{\mu(1-\mu)}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cor}(f_i, f_j) \\
&= \mu^2 + \frac{\mu(1-\mu)}{n^2}\left(\sum_{i=1}^{n}\mathrm{Cor}(f_i, f_i)\right. \\
&\qquad \left. + 2\sum_{k<l}\mathrm{Cut}(V_k, V_l) + 2\sum_{k=1}^{r}\mathrm{Cor}(V_k)\right) \\
&= \mu^2 + \frac{\mu(1-\mu)}{n} + \frac{2\mu(1-\mu)}{n^2}\sum_{k<l}\mathrm{Cut}(V_k, V_l) \\
&\qquad + \frac{2\mu(1-\mu)}{n^2}\sum_{k=1}^{r}\mathrm{Cor}(V_k). \qquad (9)
\end{aligned}$$

Therefore the expected variance of partitioned sampling can be calculated by

$$\begin{aligned}
&\mathbb{E}_M\left[\mathrm{Var}_S(\hat{f}_{part})\right] \\
&= \frac{1}{n^2}\sum_{k=1}^{r} n_k^2\mu + \frac{2}{n^2}\sum_{k<l}\left[n_k n_l \mu^2 + \mu(1-\mu)\,\mathrm{Cut}(V_k, V_l)\right] \\
&\quad - \frac{2\mu(1-\mu)}{n^2}\left(\sum_{k<l}\mathrm{Cut}(V_k, V_l) + \sum_{k=1}^{r}\mathrm{Cor}(V_k)\right) \\
&\quad - \mu^2 - \frac{\mu(1-\mu)}{n} \\
&= \frac{\mu}{n^2}\sum_{k=1}^{r} n_k^2 + \frac{2}{n^2}\sum_{k<l} n_k n_l \mu^2 - \mu^2 \\
&\quad - \frac{\mu(1-\mu)}{n} - \frac{2\mu(1-\mu)}{n^2}\sum_{k=1}^{r}\mathrm{Cor}(V_k) \\
&= \frac{\mu}{n^2}\sum_{k=1}^{r} n_k^2 - \frac{\mu^2}{n^2}\left(n^2 - 2\sum_{k<l} n_k n_l\right) - \frac{\mu(1-\mu)}{n} \\
&\quad - \frac{2\mu(1-\mu)}{n^2}\sum_{k=1}^{r}\mathrm{Cor}(V_k) \\
&= \frac{\mu}{n^2}\sum_{k=1}^{r} n_k^2 - \frac{\mu^2}{n^2}\sum_{k=1}^{r} n_k^2 - \frac{\mu(1-\mu)}{n^2}\sum_{k=1}^{r} n_k \\
&\quad - \frac{2\mu(1-\mu)}{n^2}\sum_{k=1}^{r}\mathrm{Cor}(V_k) \\
&= \frac{2\mu(1-\mu)}{n^2}\sum_{k=1}^{r}\left(\binom{n_k}{2} - \mathrm{Cor}(V_k)\right) \\
&= \frac{\mu(1-\mu)}{n^2}\sum_{k=1}^{r}\mathrm{Vol}_{G_a}(V_k) \\
&= \frac{\mu(1-\mu)}{n^2} g_r(\mathcal{P}).
\end{aligned}$$

Ignoring constant items, we need to minimize the cost function $g_r(\mathcal{P})$. □

LEMMA 5. *The Min-r-Volume problem is NP-hard to be approximated within any finite factor.*

PROOF. We establish a reduction from the r-coloring problem to the Min-r-Volume problem as follows.

Given a graph G for r-coloring. A *proper coloring* of the graph is a r-labelling to all the vertices such that the endpoints of every edge are colored differently. The r-coloring problem is to decide if there exists a proper coloring for a given graph using at most r labels. For $r \geq 3$, this problem is NP-complete [9].

Suppose that we have an approximation algorithm for Min-r-Volume on the graph G with finite approximation factor. We can use this algorithm to solve an instance of r-coloring of the graph G in polynomial time, in the following manner.

If the optimal solution of Min-r-Volume is zero, there should be no edges inside any partition. Then we color the vertices in the same partition with the same color, thus there will be no two adjacent vertices sharing the same color. This is a r-coloring of the graph G. Otherwise, if there exists a r-coloring of the graph G, we put the same colored vertices in the same partition, leading to the sum of r Volumes equal to zero. If we apply the approximation algorithm of Min-r-Volume with finite approximation factor to the graph G, we are able to distinguish whether the optimal solution of Min-r-Volume is zero or not, which indicates whether the r-coloring of the graph G exists or not.

Hence, we establish the polynomial-time reduction. □

LEMMA 6. *Let $\mathcal{P}$ be the partition produced by greedy partitioning algorithm (Algorithm 4.2) after the first iteration of all nodes. Then*

$$\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))] < \mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{naive})] + \frac{1}{4n}.$$

PROOF. The variance of naive sampling is (see Equation (7))

$$\mathrm{Var}_S(\hat{f}_{naive}) = \frac{n-r}{r(n-1)}(\bar{f} - \bar{f}^2).$$

Notice that $\mathbb{E}_M[\bar{f}] = \mu$ and

$$\mathbb{E}_M[\bar{f}^2] = \mu^2 + \frac{\mu(1-\mu)}{n} + \frac{2\mu\,(1-\mu)}{n^2}\sum_{k<l}\mathrm{Cut}(V_k, V_l) + \frac{2\mu\,(1-\mu)}{n^2}\sum_{k=1}^{r}\mathrm{Cor}(V_k).$$

which is obtained in Equation (9). Thus the expected variance of naive sampling is

$$\begin{aligned}
&\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{naive})] \\
&= \frac{(n-r)\mu(1-\mu)}{r(n-1)}\left[1 - \frac{1}{n} - \frac{2}{n^2}\sum_{i\neq j}\mathrm{Cor}(f_i, f_j)\right] \\
&= \frac{(n-r)\mu(1-\mu)}{r(n-1)n^2}\sum_{i\neq j}[1-\mathrm{Cor}(f_i, f_j)]
\end{aligned}$$

where $\sum_{i\neq j}[1-\mathrm{Cor}(f_i, f_j)]$ is the volume of graph G_a.

In the first iteration of greedy partitioning algorithm, the nodes are assigned to one of the components in the sequence $v_{s_1}, v_{s_2}, \ldots, v_{s_n}$, and the cost function is increasing during each assignment. In the k-th assignment, the increase of cost function is no more than $\sum_{l=1}^{k-1} w_{s_k s_l}/r$ where $w_{s_k s_l}$ is $1-\mathrm{Cor}(f_{s_k}, f_{s_l})$.

Thus after the first iteration, the cost function

$$g_r(\mathcal{P}) \leq \sum_{k=2}^{r}\sum_{l=1}^{k-1} w_{s_k s_l}/r < \frac{1}{r}\sum_{i\neq j}[1-\mathrm{Cor}(f_i, f_j)].$$

Therefore

$$\begin{aligned}
&\mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{part}(\mathcal{P}))] - \mathbb{E}_M[\mathrm{Var}_S(\hat{f}_{naive})] \\
&= \frac{\mu(1-\mu)}{n^2}\left[g_r(\mathcal{P}) - \frac{n-r}{r(n-1)}\sum_{i\neq j}[1-\mathrm{Cor}(f_i, f_j)]\right] \\
&< \frac{\mu(1-\mu)}{rn^2}\left(1 - \frac{n-r}{n-1}\right)\sum_{i\neq j}[1-\mathrm{Cor}(f_i, f_j)] \\
&\leq \frac{\mu(1-\mu)}{rn^2}\frac{r-1}{n-1}n(n-1) \\
&= \frac{\mu(1-\mu)(1-1/r)}{n} \\
&< \frac{1}{4n}.
\end{aligned}$$

This finishes the proof. □

B. SDP PARTITIONING ALGORITHM

In this part, we present the formulation of our SDP partitioning algorithm. The task is to find r components $\mathcal{P} = \{V_1, V_2, \ldots, V_r\}$ in order to maximize the following function:

$$g'(\mathcal{P}) = \sum_{k=1}^{r-1}\sum_{l=k+1}^{r}\mathrm{Cut}_{G_a}(V_k, V_l)$$

where $\mathrm{Cut}_{G_a}(V_k, V_l)$ is defined by $\sum_{i,j:v_i\in V_k, v_j\in V_l} w_{ij}$.

Frieze and Jerrum [8] propose an approximation algorithm for Max-r-Cut using Semi-Definite Programing (SDP) as a relaxation. We adopt their algorithm for solving our problem.

Take an equilateral simplex in $\mathbb{R}^{r-1}$ with vertices $\vec{b}_1$, $\vec{b}_2$, $\ldots$, $\vec{b}_r$. Let $\vec{c} = (\vec{b}_1 + \vec{b}_2 + \cdots + \vec{b}_r)/r$, and let $\vec{a}_k = \frac{\vec{b}_k - \vec{c}}{\|\vec{b}_k - \vec{c}\|}$ for $1 \leq k \leq r$. There is a simple property for $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_r$ that

$$\vec{a}_k \cdot \vec{a}_l = \begin{cases} 1, & \text{if } \vec{a}_k = \vec{a}_l; \\ -\frac{1}{r-1}, & \text{if } \vec{a}_k \neq \vec{a}_l. \end{cases}$$

We use $\vec{y}_i \in \{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_r\}$ to present which component node v_i is located in. If node v_i is in k-th component, then $\vec{y}_i = \vec{a}_k$. In this way, the maximization problem can be written as

$$\begin{aligned}
&\text{Maximize} && \frac{r-1}{2r}\sum_{i\neq j}[1-\mathrm{Cor}(f_i, f_j)]\,(1 - \vec{y}_i\cdot\vec{y}_j) && \text{(IP)} \\
&\text{Subject to} && \vec{y}_i \in \{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_r\}, i \in \{1, 2, \ldots, n\}.
\end{aligned}$$

To obtain the SDP relaxation, we replace $\vec{y}_i\cdot\vec{y}_j$ by (i,j)-entry of the positive semi-definite symmetric matrix Y whose diagonal elements are equal to 1, and relax $\vec{y}_i \cdot \vec{y}_j$ to be not less than $-\frac{1}{r-1}$.

$$\begin{aligned}
&\text{Maximize} && \frac{r-1}{2r}\sum_{i\neq j}[1-\mathrm{Cor}(f_i, f_j)]\,(1 - Y_{ij}) && \text{(SDP)} \\
&\text{Subject to} && Y_{ii} = 1, \forall i, \\
& && Y \succeq 0, \\
& && Y_{ij} \geq -\frac{1}{r-1}, \forall i \neq j, \\
& && Y \text{ is symmetric.}
\end{aligned}$$

Our SDP partitioning algorithm is performed by solving the above SDP problem and rounding the SDP-relaxed solution to IP-flexible solution, which is shown in Algorithm 4.1.

References

[1] S. Brakken-Thal. Gershgorinís theorem for estimating eigenvalues, 2007.

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

[3] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms*, 2001.

[4] J. T. Cox. Coalescing random walks and voter model consensus times on the torus in zd. *The Annals of Probability*, 1989.

[5] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

[6] A. Das, S. Gollapudi, R. Panigrahy, and M. Salek. Debiasing social wisdom. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, 2013.

[7] A. Dasgupta, R. Kumar, and D. Sivakumar. Social sampling. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, 2012.

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

[9] M. R. Garey and D. S. Johnson. Computer and intractability. *A Guide to the Theory of NP-Completeness*, 1979.

[10] A. Gionis, E. Terzi, and P. Tsaparas. Opinion maximization in social networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013.

[11] S. Goel, W. Mason, and D. J. Watts. Real and perceived attitude agreement in social networks. *Journal of personality and social psychology*, 2010.

[12] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

[13] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[14] Y. Li, W. Chen, Y. Wang, and Z. Zhang. Voter model on signed social networks. *Internet Mathematics*, 11(2): 93–133, 2015.

[15] T. M. Liggett. *Interacting particle systems*. Springer Science & Business Media, 2006.

[16] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 2001.

[17] J. R. Norris. *Markov chains*. Cambridge university press, 1998.

[18] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

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

[20] N. J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu, and X. Xie. We know how you live: Exploring the spectrum of urban lifestyles. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, 2013.