Partitioned Sampling of Public Opinions Based on Their Social Dynamics

Weiran Huang[1] Liang Li[2] Wei Chen[3]

[1]IIIS, Tsinghua University [2]AI Department, Ant Financial Group [3]Microsoft Research

AAAI-17 SAN FRANCISCO

How to Obtain Public Opinions?

- **Naïve Sampling:** randomly sampling a large number of individuals from the entire population and then interviewing them one by one.
  - \+ The result is unbiased. - Conducting interviews is very costly.
- **Prediction:** predicting public opinions on certain issues using historical data or online social media data which are available in the big data era.
  - \+ Such predictions cost less human effort.
  - \- They are usually biased and may lead to incorrect decisions.

Our Goal: We want to keep the estimation unbiased while saving the cost.

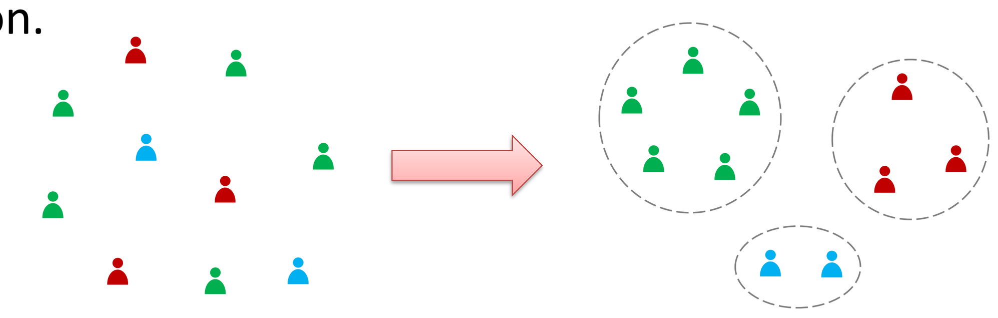
Sampling Utilizing Social Interactions

Opinion Clustering in the Real World

- People's opinions are often correlated, especially among friends in a social network, due to 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.

Main Idea

We first 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 then sample very few people in each group and aggregate the sampling results together to achieve an accurate estimation.



Formulating the OPS Problem

Given vertex set $V = \{v_1, v_2, \dots, v_n\}$, sample size budget r, opinion function $f: V \to \{0,1\}$, the task is to estimate the mean opinion $\bar{f} = \sum_{i=1}^{n} f(v_i)/n$.

- **Naïve Sampling:**
  - ➢ $\hat{f}_{naive}(V, r) = \sum_{i=1}^{r} f(x_i)/r$ where x_i is the i-th sampled node.
- **Partitioned Sampling:**
  - ➢ *Input* partition $\mathcal{P} = \{(V_1, r_1), (V_2, r_2), \dots, (V_K, r_K)\}$
  - ➢ *Output* Estimation $\hat{f}_{part}(\mathcal{P}) = \sum_{k=1}^{K} |V_k| \cdot \hat{f}_{naive}(V_k, r_k)/n$

Properties:

- Partitioned sampling is unbiased.
- Naïve sampling is a special case of partitioned sampling with $\mathcal{P} = (V, r)$.

What is a good partition? (How to evaluate the sample quality?)

- When $f(v_1), \dots, f(v_n)$ are fixed
  sample variance $\mathrm{Var}(\hat{f}_{part})$
- When $f(v_1), \dots, f(v_n)$ are random variables from a prior joint distribution
  expected sample variance $\mathbb{E}[\mathrm{Var}(\hat{f}_{part})]$

prior knowledge sampling

Optimal Partitioned Sampling (OPS) Problem

- *Input* Vertex set V, sample size budget r, pairwise opinion similarity $\sigma_{ij} = \Pr[f(v_i) = f(v_j)]$
- *Output* Optimal partition $\mathcal{P}$
- *Objective* Minimize $\mathbb{E}\left[\mathrm{Var}\left(\hat{f}_{part}(\mathcal{P})\right)\right]$

Two Tasks:

I. How to partition the vertex set V into groups?
II. How to allocate the subsample size of each group? One sample for each group!

Simple partitions: Each group only contains one sample.

Theorem 1: 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 simple partition $\mathcal{P}'$ is at least as good as partitioned sampling using the original partition $\mathcal{P}$.

Next, how to find the optimal simple partition?

Solving the OPS Problem

Assistant graph G_a: vertex set V, edge weight $w_{ij} = 1 - \sigma_{ij}$.

For a simple partition $\mathcal{P} = \{(V_1, 1), \dots, (V_r, 1)\}$ of V, each group V_k's volume in G_a is $Vol_{G_a}(V_k) = \sum_{v_i, v_j \in V_k} w_{ij}$.

Theorem 2: The optimal simple partition minimizes the sum of all groups' volumes in G_a. Moreover, for any simple partition $\mathcal{P} = \{(V_1, 1), \dots, (V_r, 1)\}$ of V,

$$\mathbb{E}\left[\mathrm{Var}\left(\hat{f}_{part}(\mathcal{P})\right)\right] = \frac{\sum_{k=1}^{r} Vol_{G_a}(V_k)}{2n^2}.$$

➢ Good partition → small volumes → small weights in each group → large opinion similarities in each group → people holding similar opinions are in the same group

Algorithms for Min-r-Partition

- SDP partitioning (adapted from SDP for Max-r-Cut (Frieze and Jerrum 1997))
  - Too slow and infeasible to run on large graph
- Greedy partitioning: greedily put the ungrouped nodes into the group such that the objective (sum of all groups' volumes) increase the least, and then repeat the greedy assignment iteratively.

Proposition 1: Partitioned sampling using the simple partition generated by the greedy partitioning algorithm is at least as good as naïve sampling.

When opinion similarities are inaccurate or even missing …

a) Force the partition to be balanced (all groups have the same size) in the Greedy partitioning.
   Theorem 3: Partitioned sampling using any *balanced* simple partition is at least as good as naïve sampling.
b) To test the robustness of the partitioned sampling method, in the experiment on the real-world network, we input the following perturbed opinion similarities to test:
   - all the opinion similarity information between disconnected nodes is removed
   - the rest opinion similarities are perturbed more than 30%

   Observation: The performance using perturbed inputs is very close to the performance using exact inputs.
c) Using an opinion evolution model to characterize social interactions, and extracting the similarities from the model. The model essentially provides a more compact representation than pairwise similarities.

Voter Model with Innate Opinions (VIO Model)

Social graph $G = (V, A)$: A_{ij} represents the opinion influence from person v_j to v_i.

Each node v_i at any time t is associated with both

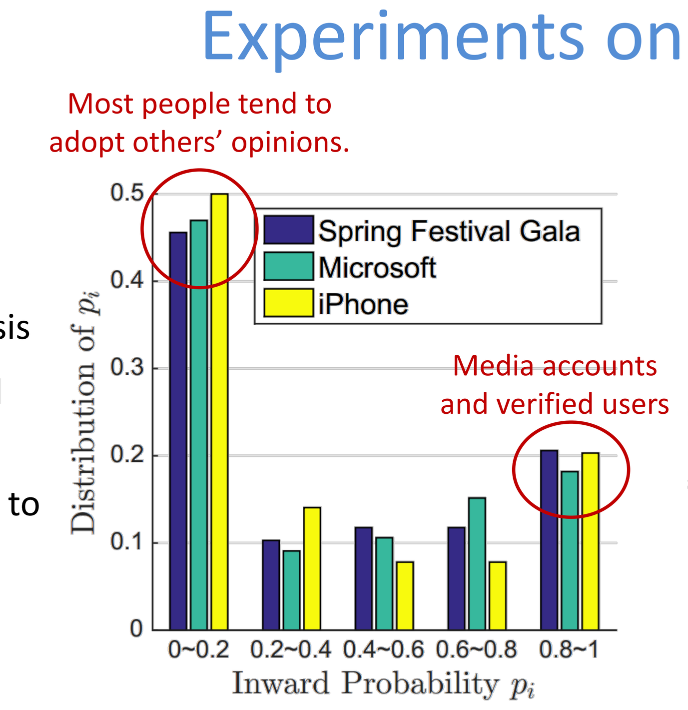
- an innate opinion $f^{(0)}(v_i)$: generated from an i.i.d. Bernoulli distribution, and unchanged from external influences.
- an expressed opinion $f^{(t)}(v_i)$: shaped by the opinions of its neighbors, and the one obtained by sampling.

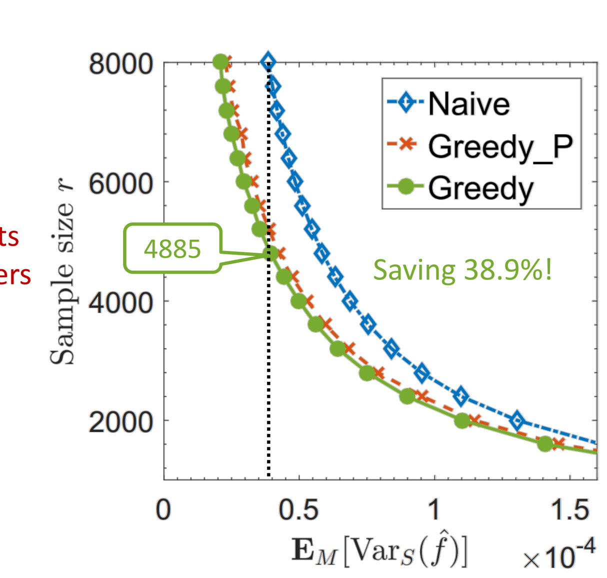
Each node v_i updates its expressed opinion according to its **Poisson process with rate** λ_i independently:

1. set to its innate opinion $f^{(t)}(v_i) = f^{(0)}(v_i)$ with **inward probability** p_i,
2. with probability $1 - p_i$, randomly selects one of its out-neighbors v_j with probability proportional to the weight of edge (v_i, v_j), and sets its expressed opinion to v_j's expressed opinion $f^{(t)}(v_i) = f^{(t)}(v_j)$.

Experiments on Weibo Dataset

- Select the topics
- Sentiment analysis
- Collect the social relationship
- Using VIO model to fit the data





- 40,787 nodes and 165,956 directed edges
- Set λ_i to the number of v_i's tweets in a year
- Using both exact and perturbed opinion similarities as inputs