Opinion Clustering in the Real World

which can be constructed efficiently, such that partitioned sampling using the simple Theorem 1:

\[ \text{Partitioned sampling is unbiased.} \]

\[ \mathbb{E} \left[ \| f_{\text{part}}(\mathcal{P}) \| \right] = \sum_{i=1}^{n} \hat{V}_{\text{true}}(V_i) \]

\[ \frac{1}{2n^2} \]

• 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) increases 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 naive 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.

\[ \text{Theorem 3: Partitioned sampling using any balanced simple partition is at least as good as naive sampling.} \]

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 similarity 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 \( v_i \) to \( v_j \).

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 \( \lambda_i \) independently:

1. set to its innate opinion \( f^{(0)}(v_i) \) 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 \( f^{(t)}(v_j) = f^{(t)}(v_i) \).

Experiments on Weibo Dataset

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

Most people tend to adopt others’ opinions.

40,787 nodes and 165,956 directed edges

Set \( \lambda_i \) to the number of \( v_i \)’s tweets in a year

Using both exact and perturbed opinion similarities as inputs

Solving the OPS Problem

Assistant graph \( G_a \): vertex set \( V, \) edge weight \( w_{ij} = 1 - \sigma_{ij} \). For a simple partition \( \mathcal{P} = \{(V_1), \ldots, (V_r)\} \) of \( V \), each group \( V_k \)'s volume in \( G_a \) is \( \hat{V}_{\text{true}}(V_k) = \sum_{j \in V_k} w_{ij} \).

Theorem 2: The optimal simple partition minimizes the sum of all groups’ volumes in \( G_a \).

Moreover, for any simple partition \( \mathcal{P} = \{(V_1), \ldots, (V_r)\} \) of \( V \),

\[ \mathbb{E} \left[ \text{Var} \left( f_{\text{part}}(\mathcal{P}) \right) \right] = \frac{\sum_{k=1}^{r} \hat{V}_{\text{true}}(V_k)}{2n^2}. \]

Properties:

- Partitioned sampling is unbiased.
- Naive 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 \( f(v_1), \ldots, f(v_n) \) are fixed
  sample variance \( \text{Var}(f_{\text{part}}) \)

- When \( f(v_1), \ldots, f(v_n) \) are random variables from a prior joint distribution
  expected sample variance \( \mathbb{E} \left[ \text{Var}(f_{\text{part}}) \right] \)

Optimal Partitioned Sampling (OPS) Problem

Input

- Vertex set \( V, \) sample size budget \( r \), pairwise opinion similarity \( \sigma_{ij} = \text{Pr}[f(v_i) = f(v_j)] \)

Output

- Optimal partition \( \mathcal{P} \)

Objective

- Minimize \( \mathbb{E} \left[ \text{Var}(f_{\text{part}}(\mathcal{P})) \right] \)

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?