Adversaries with Limited Information in the Friedkin-Johnsen Model

SIJING TU · STEFAN NEUMANN · ARISTIDES GIONIS
{sijing, neum, argioni}@kth.se, KTH Royal Institute of Technology, Sweden

MOTIVATION

Phenomenon:

- Russian military and intelligence services have been using online social networks to sow discord and discredit legitimate political institutions.
- A recent analysis regarding the Iranian disinformation campaigns shows that their main goal is to pit groups against each other.

Observation:

- The network structure is easier to obtain compared to users' opinions.

Research Question:

- How much additional discord can attackers instigate in online social networks, given only the network structure?

OPINION FORMATION AND NETWORK DISCORD

Let $G = (V, E, w)$ be a weighted undirected graph.

Opinion Formation: Friedkin–Johnsen model [2]

Each user $u \in V$ has

- an expressed opinion $z_u \in [-1, 1]$, which depends on the network and which changes over time due to peer pressure,
- an innate opinion $s_u \in [-1, 1]$ that is fixed.

The expressed opinions are updated based on the update rule:

$$\mathbf{z}_u^{(t+1)} = \frac{s_u + \sum_{(u,v)\in E} w_{u,v} z_v^{(t)}}{1 + \sum_{v\in N(u)} w_{u,v}}$$

Equilibrium opinions for $t \to \infty$

$$\mathbf{z}^* = (\mathbf{I} + \mathbf{L})^{-1}\mathbf{s},$$

where $\mathbf{L}$ is the graph Laplacian and $\mathbf{I}$ is the identity matrix.

Network Discord:

Name	Notation	Matrix
Polarization	$\mathcal{P}(\mathbf{L})$	$(\mathbf{I}+\mathbf{L})^{-1}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\intercal}{n})(\mathbf{I}+\mathbf{L})^{-1}$
Disagreement	$\mathcal{D}(\mathbf{L})$	$(\mathbf{L}+\mathbf{I})^{-1}\mathbf{L}(\mathbf{L}+\mathbf{I})^{-1}$

Discord matrix $A(\mathbf{L}) \in \{\mathcal{P}(\mathbf{L}), \mathcal{D}(\mathbf{L})\}$:

- Polarization $\mathcal{P}_{G,\mathbf{s}}$
  measures the variance of the expressed opinions:
  $\mathcal{P}_{G,\mathbf{s}} = \sum_{v\in V}(z_v - \bar{\mathbf{z}})^2 = \mathbf{s}^\intercal \mathcal{P}(\mathbf{L})\,\mathbf{s}$.
- Disagreement $\mathcal{D}_{G,\mathbf{s}}$
  measures the differences between the expressed opinions:
  $\mathcal{D}_{G,\mathbf{s}} = \sum_{(u,v)\in E} w_{u,v}(z_u - z_v)^2 = \mathbf{s}^\intercal \mathcal{D}(\mathbf{L})\,\mathbf{s}$.

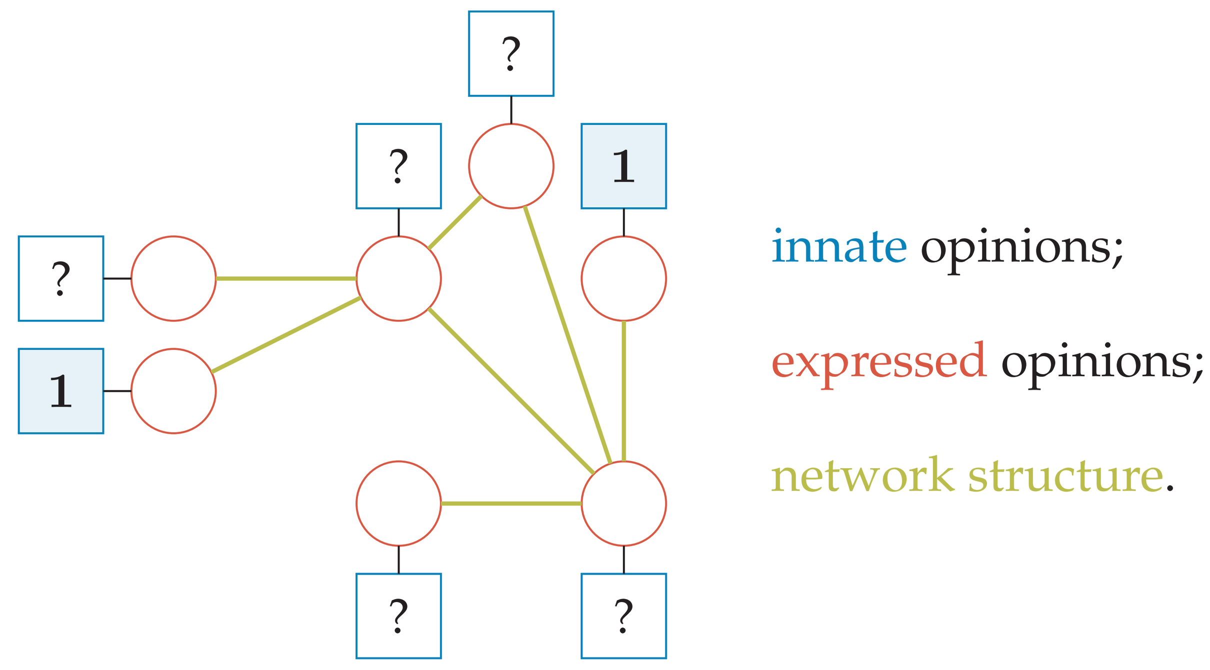
Problem (Maximizing Discord with Full Information [1, 3]). Radicalize k users' innate opinions by setting their innate opinions to 1.

$$\begin{aligned} &\max_{\mathbf{s}} && \mathbf{s}^\intercal A(\mathbf{L})\,\mathbf{s}, \\ &\text{such that} && \|\mathbf{s} - \mathbf{s}_0\|_0 = k, \text{ and} \\ &&& \mathbf{s}(u) \in \{\mathbf{s}_0(u), 1\} \text{ for all } u \in V. \end{aligned}$$

ADVERSARIES WITH LIMITED INFORMATION

We assume a weak **adversary with limited-information** that

- has access to the network structure;
- but **does not** have access to the innate opinions;
- and can radicalize k nodes' innate opinions.



LIMITED-INFORMATION MODEL

Observation: Assume that the innate opinions are centered around some constant, the adversary applies the following strategy:

- It pretends the initial innate opinions of all the nodes are 0;
- it finds the nodes that maximize the discord in this simplified setting;
- it radicalizes these selected nodes in this simplified model in the original problem.

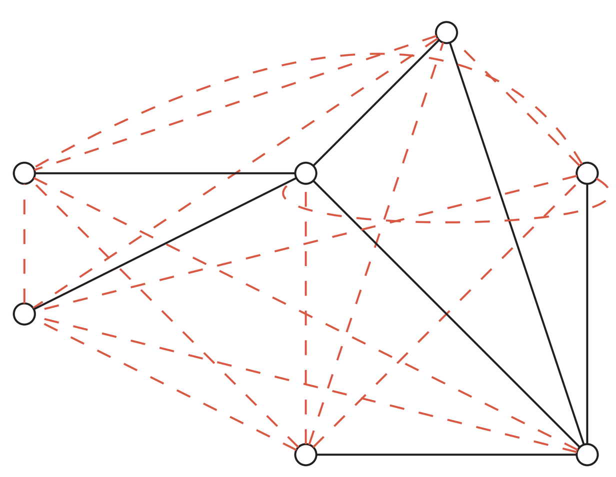
Problem (Maximizing Discord with Limited Information).

$$\begin{aligned} &\max_{\mathbf{s}} && \mathbf{s}^\intercal A(\mathbf{L})\,\mathbf{s}, \\ &s.t. && \|\mathbf{s} - \mathbf{0}\|_0 = k, \text{ and} \\ &&& \mathbf{s} \in \{0, 1\}^n. \end{aligned}$$

Connection: When the innate opinions have small variance, and other mild assumptions hold, **any $\mathcal{O}(1)$-approximate solution to the limited-information problem is a $\mathcal{O}(1)$-approximation solution to the full-information problem**.

Analysis: Solving the above limited-information problem is equivalent to solving a constrained Max-Cut problem with positive and *negative* edge weights.

- We apply a semidefinite-relaxation based algorithm to solve it.
- We compare our algorithm with greedy algorithms and other heuristics.

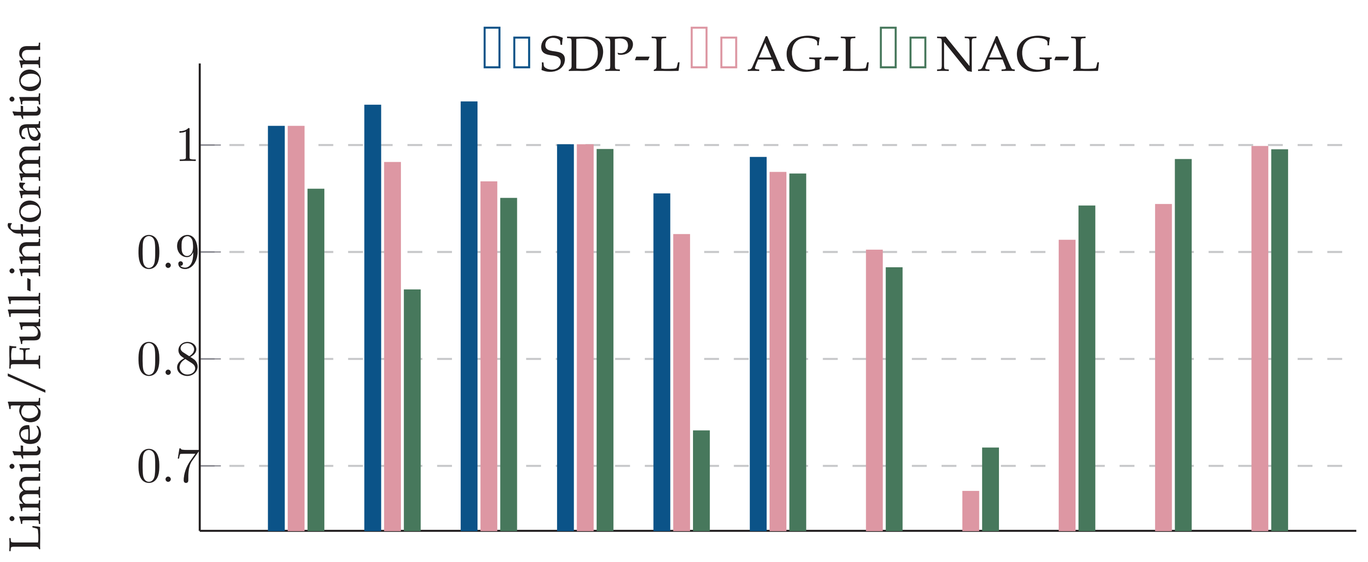


- The negative edges are in red,
- positive edges are in black.
- We partition the nodes into subsets of sizes $(n-k, k)$ to maximize the cut.

- The problem is **NP**-hard;
- The problem has constant approximation ratio when $k = \Omega(n)$.

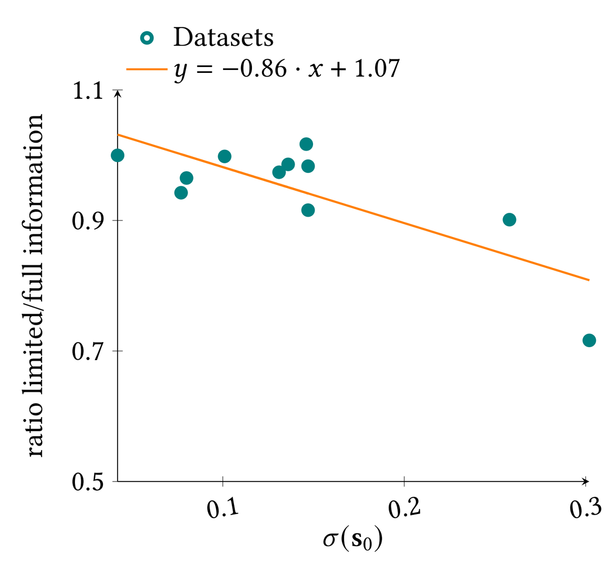
EXPERIMENTS

Results on all datasets: SDP-L is the best among limited-information algorithms, limited-information algorithms are at most a factor of 1.4 worse. (SDP-L: SDP-based; AG-L: Adaptive-Greedy; NAG-L: NonAdaptive-Greedy)

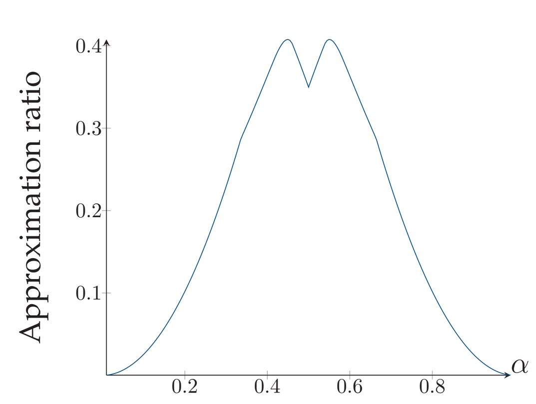


All the datasets (sorted according to n, SDP-L runs on 6 smaller datasets)

Note: Baselines such as selecting high-degree nodes perform much worse.



(a) standard deviation of opinions, $R^2 = 0.62$



(b) approximation ratio

[1] M. F. Chen and M. Z. Racz. An adversarial model of network disruption: Maximizing disagreement and polarization in social networks. *IEEE Transactions on Network Science and Engineering*, 2021.
[2] N. E. Friedkin and E. C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 1990.
[3] J. Gaitonde, J. M. Kleinberg, and É. Tardos. Adversarial perturbations of opinion dynamics in networks. In *EC*, 2020.