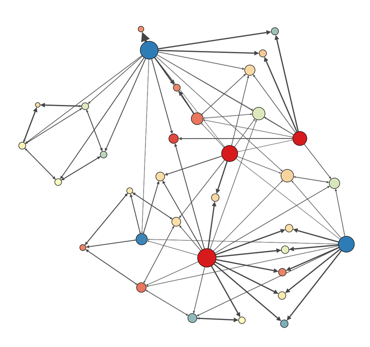
A Distance Measure for the Analysis of Polar Opinion Dynamics in Social Networks

Victor Amelkin*, Petko Bogdanov†, Ambuj K. Singh*

*University of California, Santa Barbara †University at Albany-SUNY

Introduction



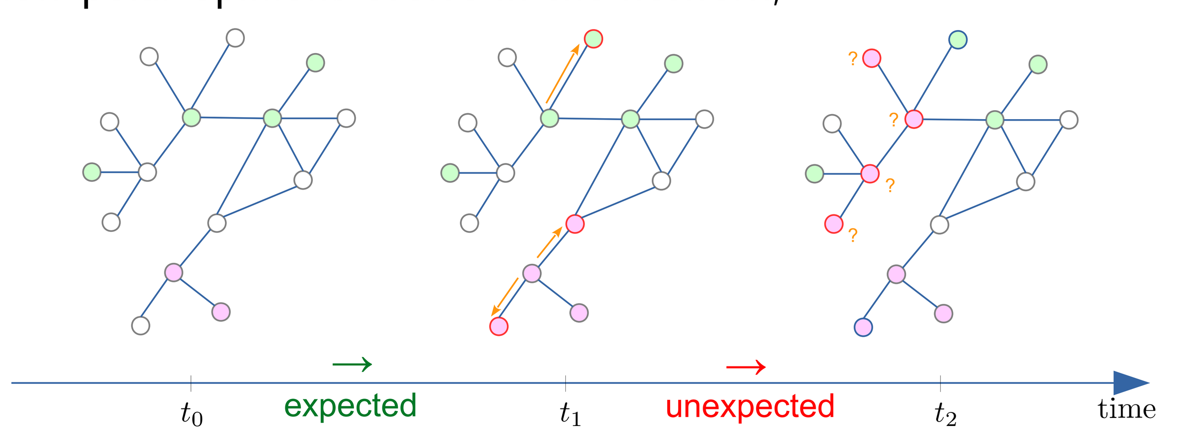
- Directed sparse social network, $|V| = n$, $|E| = m$
- Opinions are *polar* (e.g., 🐴 vs. 🐘)$\in \{+1, 0, -1\}$
- **Network state** G_t – opinions of all users at time t
- Network structure does not change significantly
- User opinions evolve

Q1: When does the network "behave" unexpectedly?

Q2: What will the future opinions of select users be?

Problem

- How to quantify the distance $d(G_1, G_2)$ between network states, so that the distance measure $d(\bullet, \bullet)$
  - ▷ captures how polar opinions evolve in the network;



  - ▷ is efficiently computable (applicable to large-scale networks) and metric.

Earth Mover's Distance as a Core Primitive

- Earth Mover's Distance (EMD) – "edit distance for histograms" [1]

$$\mathrm{EMD}(P,Q,D) = \sum_{i,j=1}^{n} D_{ij}\hat{f}_{ij} \Big/ \sum_{i,j=1}^{n} \hat{f}_{ij},$$

$$\sum_{i,j=1}^{n} f_{ij}D_{ij} \to \min,\ \sum_{i,j=1}^{n} f_{ij} = \min\left\{\sum_{i=1}^{n} P_i, \sum_{i=1}^{n} Q_i\right\}$$

$$f_{ij} \ge 0, \sum_{j=1}^{n} f_{ij} \le P_i, \sum_{i=1}^{n} f_{ij} \le Q_j, (1 \le i,j \le n)$$

- Extendable (EMD* [2]) to histograms derived from network states

References

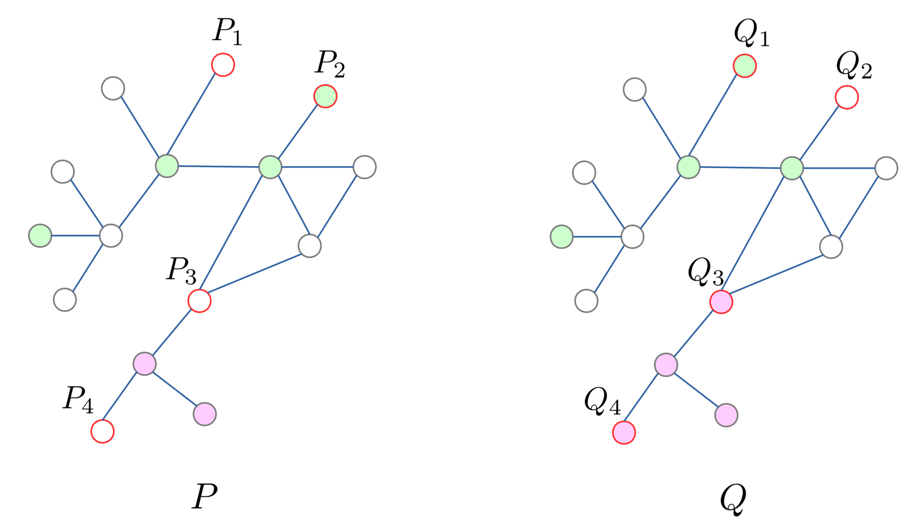
[1] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[2] —, "A distance measure for the analysis of polar opinion dynamics in social networks (Full Paper)," *available at* http://cs.ucsb.edu/~victor/pub/ucsb/dbl/snd/snd-full.html.

[3] R. K. Ahuja, K. Mehlhorn, J. Orlin, and R. E. Tarjan, "Faster algorithms for the shortest path problem," *Journal of the ACM*, vol. 37, no. 2, pp. 213–223, 1990.

[4] R. K. Ahuja, J. B. Orlin, C. Stein, and R. E. Tarjan, "Improved algorithms for bipartite network flow," *SIAM Journal on Computing*, vol. 23, no. 5, pp. 906–933, 1994.

Social Network Distance (SND) – Intuition

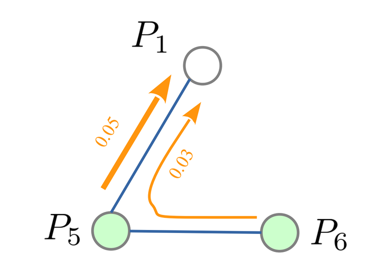


$$\mathrm{SND}(P,Q) \approx -\log \mathbb{P}\left\{\begin{matrix} P_1 \bigcirc \rightsquigarrow \bigcirc Q_1,\ P_3 \bigcirc \rightsquigarrow \bigcirc Q_3, \\ P_2 \bigcirc \rightsquigarrow \bigcirc Q_2,\ P_4 \bigcirc \rightsquigarrow \bigcirc Q_4. \end{matrix}\right\}$$

- Exact computation of $\mathbb{P}$ is computationally infeasible
- Assume that user activations are independent

$$\mathbb{P}\{P_1 \bigcirc \rightsquigarrow \bigcirc Q_1 \mid P_3 \bigcirc \rightsquigarrow \bigcirc Q_3\} = \mathbb{P}\{P_1 \bigcirc \rightsquigarrow \bigcirc Q_1\}$$

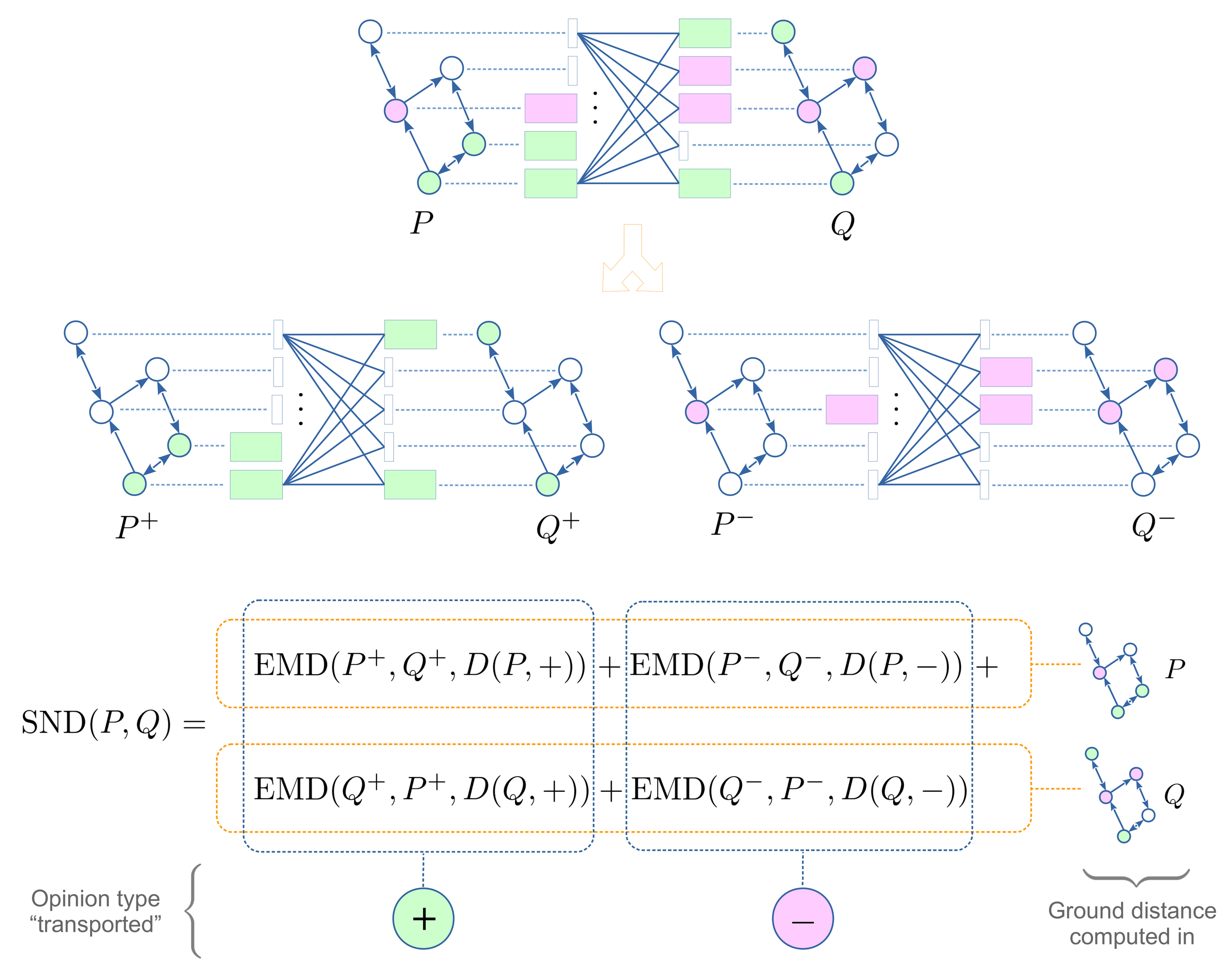
- Assume that opinions spread (and get adopted) via most likely paths



$$\mathbb{P}\left\{P_5 \xrightarrow{infects} P_1\right\} = 0.05 > 0.03 = \mathbb{P}\left\{P_6 \xrightarrow{infects} P_1\right\}$$

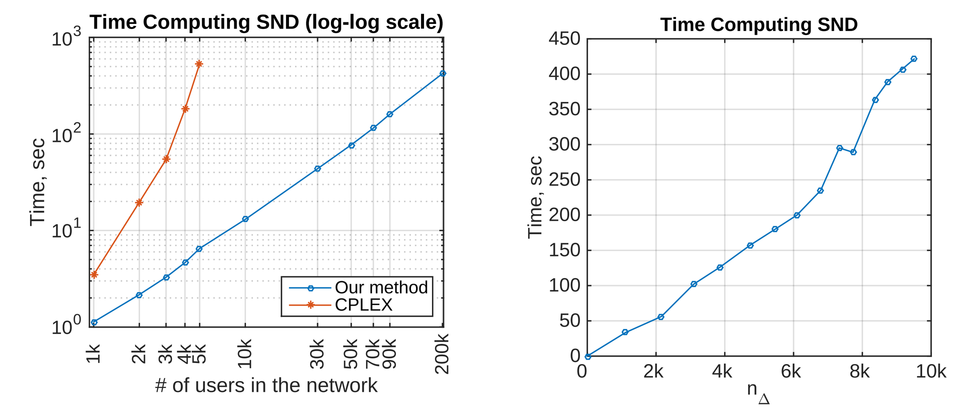
- As a result, SND can be defined as **a transportation problem**

Social Network Distance (SND) – Definition



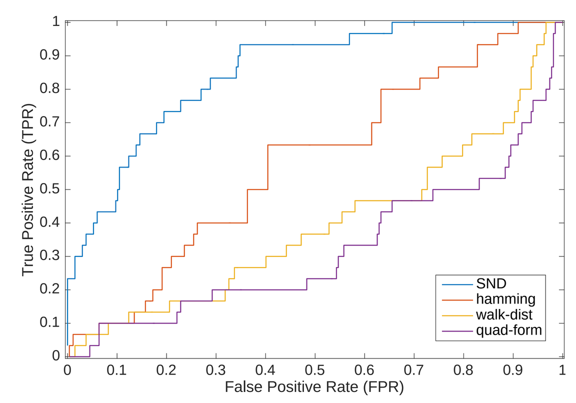
Efficient Computation of SND

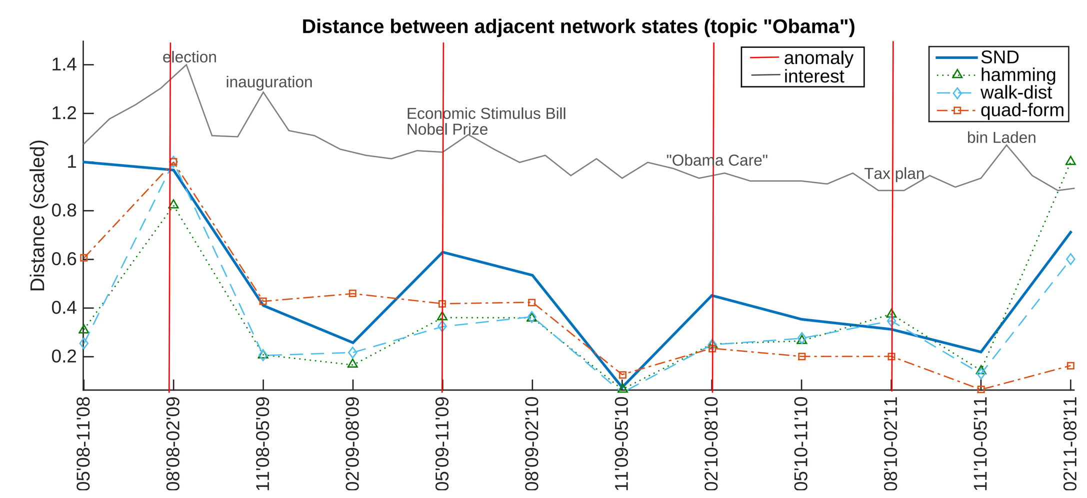
- Direct computation of $\mathrm{EMD}(P,Q,D)$ (and, $\mathrm{SND}(P,Q,D)$) over sparse network involves computing all-to-all shortest paths ($\mathcal{O}(n^2 \log n)$) and solving a transportation problem ($\mathcal{O}(n^3 \log n)$).
- Key ideas for computing exact $\mathrm{EMD}(P,Q,D)$ in pseudo-linear time:
  - ▷ Assume number n_Δ of users who changed their opinions $\ll n$, and $D_{ij} \in \mathbb{Z}^+ < U = const$.
  - ▷ Reduce the optimization problem using semi-metricity of D in SND.
  - ▷ Efficiently compute D (few-to-all shortest paths) via Dijkstra with radix and Fibonacci heaps [3]; and the underlying transportation problem (unbalanced min-cost flow) via modified Goldberg-Tarjan algorithm [4].
- **Time complexity:** $\mathcal{O}(n_\Delta(n \log \sqrt{U} + n_\Delta^2 \log(n_\Delta nU))) = \boldsymbol{\mathcal{O}(n)}$



Application I – Anomaly Detection

- Anomalies—spikes in the series of adjacent network states distances.
- Application to synthetic and Twitter data:





- SND usually spikes during "polarizing events"

Application II – User Opinion Prediction

- Predicted opinions make the distance to the current network state as close to the estimate as possible.

User Opinion Prediction Accuracy, %

Method	Synthetic Data		Twitter Data	
	μ	σ	μ	σ
SND	**74.33**	**2.65**	**75.63**	**5.60**
hamming	68.44	12.34	68.13	5.80
quad-form	66.67	13.58	67.50	9.63
walk-dist	56.22	15.35	31.88	9.98
icc-simulation	**76.25**	**9.54**	59.38	4.17
ltc-simulation	67.50	11.65	58.75	5.18
icc-max-likelihood	67.41	7.03	57.50	8.02
ltc-max-likelihood	57.50	8.45	55.63	11.78
community-lp	65.25	9.43	56.87	8.43

Supported by, or in part by, the U.S. Army Research Laboratory and the U.S. Army Research Office under grant number W911NF-15-1-0577.

{victor,ambuj}@cs.ucsb.edu, pbogdanov@albany.edu