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

VICTOR AMELKIN*, University of Pennsylvania
PETKO BOGDANOV, University at Albany—SUNY
AMBUJ K. SINGH, University of California, Santa Barbara

Analysis of opinion dynamics in social networks plays an important role in today's life. For predicting users' political preference, it is particularly important to be able to analyze the dynamics of competing polar opinions, such as pro-Democrat vs. pro-Republican. *While observing the evolution of polar opinions in a social network over time, can we tell when the network evolved abnormally? Furthermore, can we predict how the opinions of the users will change in the future?* To answer such questions, it is insufficient to study individual user behavior, since opinions can spread beyond users' ego-networks. Instead, we need to consider the opinion dynamics of all users simultaneously and capture the connection between the individuals' behavior and the global evolution pattern of the social network.

In this work, we introduce the Social Network Distance (SND)—a distance measure that quantifies the likelihood of evolution of one snapshot of a social network into another snapshot under a chosen model of polar opinion dynamics. SND has a rich semantics of a transportation problem, yet, is computable in time linear in the number of users and, as such, is applicable to large-scale online social networks. In our experiments with synthetic and Twitter data, we demonstrate the utility of our distance measure for anomalous event detection. It achieves a true positive rate of 0.83, twice as high as that of alternatives. The same predictions presented in precision-recall space show that SND retains perfect precision for recall up to 0.2. Its precision then decreases while maintaining more than 2-fold improvement over alternatives for recall up to 0.95. When used for opinion prediction in Twitter data, SND's accuracy is 75.6%, which is 7.5% higher than that of the next best method.

CCS Concepts: • **Mathematics of computing → Graph algorithms**; Time series analysis; • **Theory of computation → Graph algorithms analysis**; • **Information systems → Social networks**; **Similarity measures**; **Data mining**; • **Human-centered computing** → *Social networks*; *Social network analysis*;

Additional Key Words and Phrases: social network; opinion dynamics; competing opinions; polar opinions; polarization; distance measure; time-series; anomaly detection; opinion prediction; model-driven analysis; Earth Mover's Distance; Wasserstein metric; transportation problem; minimum-cost network flow.

ACM Reference format:
Victor Amelkin, Petko Bogdanov, and Ambuj K. Singh. 2019. A Distance Measure for the Analysis of Polar Opinion Dynamics in Social Networks. *ACM Trans. Knowl. Discov. Data.* 1, 1, Article 1 (January 2019), 34 pages.
https://doi.org/10.1145/3332168

* The bulk of this work was completed when the author was with the University of California, Santa Barbara.

1556-4681/2019/1-ART1 $15.00
https://doi.org/10.1145/3332168

1 INTRODUCTION

Analysis of opinion formation in the society plays an important role in today's life. Businesses are interested in advertising their products in social networks relying on viral marketing. Political strategists are interested in predicting an election outcome based on the observed sentiment change of a sample of voters. Mass media and security analysts may be interested in a timely discovery of anomalies based on how a social network "behaves". Thus, it is important to enable modeling and prediction of user opinion evolution in a social network.

How can we quantify the change in opinions of users with respect to their expected behavior in a social network? How can we predict how the opinions of individual users will evolve in the future? Having observed the evolution of user opinions over time, can we tell when the opinions evolved abnormally? To answer such questions, we need a distance measure for the comparison of states of a social network that explicitly models user opinion evolution, incorporating both the distribution of user opinions at two time instances and the network structure that defines the pathways for opinion dissemination. In this work, we develop such a distance measure and employ it for anomaly detection and opinion prediction.

While the dynamics of a social network can be characterized by evolution of both the network's structure and user opinions, in this paper we focus on the latter. We assume that there are two *polar opinions* in the network, *positive* and *negative*. Users having no or an unknown opinion are termed *neutral*, while those expressing opinion—*active*. A *network state* is comprised of the opinions of all network users at a given time. Polar opinions *compete* in that users are less willing to spread opinions different from their own, yet, are more eager to spread "friendly" opinions. Such competition may arise when the notions the opinions relate to, such as political parties or smartphone brands, are inherently competing.

Having observed the behavior of a social network's users over time and quantified their opinions, we obtain a time series of network states. Its analysis—whether anomaly detection or future state forecasting—is, however, problematic, as network states do not naturally belong to any vector space, and the numerous existing time series analysis techniques cannot be readily applied. Our approach is to treat network states as members of a metric space induced by a distance measure governed by both the network's structure and user opinions. We propose a semantically and mathematically appealing, as well as efficiently computable distance measure *Social Network Distance (SND)* for the social network states containing polar opinions and demonstrate its utility in two applications. First, we detect which network states in a series are anomalous with respect to the expected opinion evolution, where the latter is determined by a chosen model of polar opinion dynamics. Second, we predict unknown opinions of individual users in a partially observed network state based on the historical dynamics of other users' opinions.

In this work, we make the following specific *contributions*:

- We propose SND—the first distance measure suitable for the comparison of social network states containing polar opinions under a chosen model of opinion dynamics.
- We develop a scalable method for exact computation of SND in time linear in the number of network users. This is achieved via exploiting the special structure of the transportation problem underlying SND and the use of special shortest path and minimum-cost network flow algorithms.
- We demonstrate the utility of SND in two applications with both synthetic and Twitter data. Using SND for anomaly detection, we achieve a true positive rate of 0.83, twice as high as that of alternatives. The same predictions presented in precision-recall space show that SND retains perfect precision for recall up to 0.2. Its precision then decreases while maintaining

more than 2-fold improvement over alternatives for recall up to 0.95. When used for user opinion in Twitter data, SND's prediction accuracy is 75.6%, which is 7.5% higher than that of the next best method.

2 PRELIMINARIES

$\mathbb{1}$	vectors of all ones
$\mathbb{0}$	vector of all zeroes
$\otimes$	Kronecker product
$\text{diag}(v)$	diagonal matrix with vector v as the main diagonal
$G(V, E)$	network with nodes V ($\|V\| = n$) and edges E ($\|V\| = m$)
$X, X(t) \in \{+1, 0, -1\}^n$	network state (at time t) comprised of all users' opinions
$X_i, X_i(t) \in \{+1, 0, -1\}$	opinion of user i in network state X (at time t)
I_P^v	set of users holding opinion v in network state P
$\mathbb{P}_{ij}(P, v)$	likelihood of user j acquiring opinion v from user i in network state P
n_Δ	number of users who changed opinion between two network states

Table 1. Notation summary

2.1 Network and network states

We are given a social network $G(V, E)$, where V ($|V| = n$) is the set of nodes (users) and E is the set of edges (social ties). At each point in time, each user holds a quantified opinion on the chosen topic of interest. In this work, we will quantify the opinions on a discrete scale $\{+1, 0, -1\}$[1], with 0 standing for neutrality, and $+1$ and -1 corresponding to two polar alternatives, such as the Democrats vs. the Republicans or iOS vs Android. Note, however, that the algorithmic results we obtain in this work also hold for a more general case of any finite discrete opinion scale. The opinions of all network users at time t comprise the *network state* $G(t) \in \{+1, 0, -1\}^n$ at time t, where $G_i(t) \in \{+1, 0, -1\}$ is the opinion of user i.

2.2 User opinions and their dynamics

We assume that the dynamics of user opinions is governed by a learned in advance opinion dynamics model $\mathcal{M}$ that provides $\mathbb{P}_{ij}(P, v)$—the likelihood of user j adopting opinion $P_i = v \in \{+1, -1\}$ of user i in network state P. In Appendix A.1, we provide two examples of definitions of $\mathbb{P}_{ij}(P, v)$ for the variants of Independent Cascade [11] and Linear Threshold [9] models supporting competing opinions. In our experiments, however, we will assume a simple and intuitive opinion dynamics model, using the following definition of (log-)likelihoods $\mathbb{P}_{ij}(P, v)$:

$$-\log \mathbb{P}_{ij}(P, v) = \begin{cases} c_{adverse} & \text{if } P_i = -v \text{ or } P_j = -v, \\ c_{neutral} & \text{else if } P_i = 0, \\ c_{friendly} & \text{else if } P_i = v \text{ and } P_j \neq -v, \end{cases} \tag{1}$$

where $c_{adverse}, c_{neutral}, c_{friendly} \in \mathbb{R}^+$ are constant log-likelihoods of adopting adverse, neutral, or friendly opinion (relatively to opinion v), respectively. Thus, users willingly spread opinions similar to their own ($c_{friendly}$ is small), are unwilling to spread adverse opinions ($c_{adverse}$ is large), with the behavior of neutral users being somewhere in-between ($c_{friendly} < c_{neutral} < c_{adverse}$).

[1] There is a great body of research on the methods for opinion classification based on user-generated content, including [49, 51]. Our focus is, however, on the analysis of *how opinions spread*, rather than on how to quantify them.

2.3 Earth Mover's Distance

In this work, we will target computing distances between network states $P, Q \in \{+1, 0, -1\}^n$ using a distance measure, whose semantics will be similar to that of one well-studied distance measure—Earth Mover's Distance (EMD) [46]. Originally defined as an edit-distance for histograms (which, in our case, can be seen simply as n-dimensional non-negative vectors), EMD computes the cost of an optimal transformation of one histogram into another, where an elementary edit operation is transportation of a unit of mass from one histogram bin to another bin. The costs of these elementary transforms are collectively referred to as the *ground distance*.

Formally, EMD between two histograms $P \in \mathbb{R}^{+n}$ and $Q \in \mathbb{R}^{+m}$ (that, in our case, will be derived from network states) over ground distance $D \in \mathbb{R}^{+n \times m}$ (that, in our case, will be defined based on the distances between users in the network and the likelihoods of the opinions spreading between them) is the solution to the problem of optimal mass transportation from suppliers $\{P_i\}$ to consumers $\{Q_j\}$ with respect to transportation costs $\{D_{ij}\}$:

$$\mathrm{EMD}(P, Q, D) = \sum_{i=1}^{n}\sum_{j=1}^{m} D_{ij}\widehat{f}_{ij} \;/\; \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{f}_{ij}, \tag{2}$$

$$\{\widehat{f}_{ij}\} = \arg\min_{\{f_{ij}\}} \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}D_{ij},\ \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = \min\Big\{\sum_{i=1}^{n} P_i, \sum_{j=1}^{m} Q_j\Big\},$$

$$f_{ij} \geq 0,\ \sum_{j=1}^{m} f_{ij} \leq P_i,\ \sum_{i=1}^{n} f_{ij} \leq Q_j,\ (1 \leq i \leq n, 1 \leq j \leq m),$$

where $\{\widehat{f}_{ij}\}_{n \times m}$ is an optimal solution or transportation plan.

In addition to having the semantics that will suit our distance measure design goals, EMD is metric, as the following theorem states.

Theorem 2.1 (Metricity of EMD [46]). *If all network states under comparison have equal total masses, and the underlying ground distance is metric, then EMD is metric.*

Metricity of EMD will allow our own EMD-based distance measure to also be metric, which—in addition to making a distance measure "natural"—can be exploited to improve practical performance of distance measure-based algorithms in applications [13].

3 DISTANCE MEASURE FOR NETWORK STATES WITH POLAR OPINIONS

The *central problem* we address in this paper is as follows:

Problem. *Given two network states $P, Q \in \{+1, 0, -1\}^n$ and assuming that the dynamics of user opinions is governed by model $\mathcal{M}$, define and compute the distance between network states P and Q reflecting how likely one of these network states has evolved into another under model $\mathcal{M}$.*

According to the problem definition above, if network state Q has evolved from network state P closely following model $\mathcal{M}$, then the resulting distance should be small; if, however, either network states P and Q are unrelated, or they are related, but the opinions evolved following a model very different from $\mathcal{M}$, then the distance should be large. The existing distance measures, including vector space (ℓ_p-like), graph-based, iterative, and feature-based ones—reviewed in detail in Sec. 7.2—do not possess such a semantics, and we need to look for a new more suitable distance measure.

The model $\mathcal{M}$ that we assume in the problem's definition to be known can be learned from data. For example, we can pick multiple general models, such as Independent Cascade or Linear

Threshold models, fit them to data [22], and choose the one that fits the data best. Alternatively, the model may come from domain knowledge.

Next, we will, first, translate the above defined problem into the formal language of probability, and, after that, address the question of making the obtained formalization tractable.

Let us put $X(t)$ to be the state of the network at time t, and $X(1,\ldots,k) = X(1),\ldots,X(k)$ to be a *network state evolution path*—a sequence of states over which the network evolved. Then, the likelihood of $X(1,\ldots,k)$ is defined as

$$\mathbb{P}\{X(1,\ldots,k)\} = \prod_{t=2}^{k} \mathbb{P}\{X(t) \mid X(t-1)\}.$$

Using this notation, a *perfect distance measure* can be defined as

$$d_{\mathcal{M}}(P,Q) = \sum_{k\geq 2} \sum_{\substack{X(1,\ldots,k) \\ X(1)=P, X(k)=Q}} \mathbb{P}\{X(1,\ldots,k)\}. \tag{3}$$

According to its definition, $d_{\mathcal{M}}(P,Q)$ measures the likelihood of user opinions evolving from state $X(1) = P$ to state $X(k) = Q$ along all possible network state evolution paths—as illustrated in Fig. 1—where the opinion change likelihoods are determined by the underlying model $\mathcal{M}$.

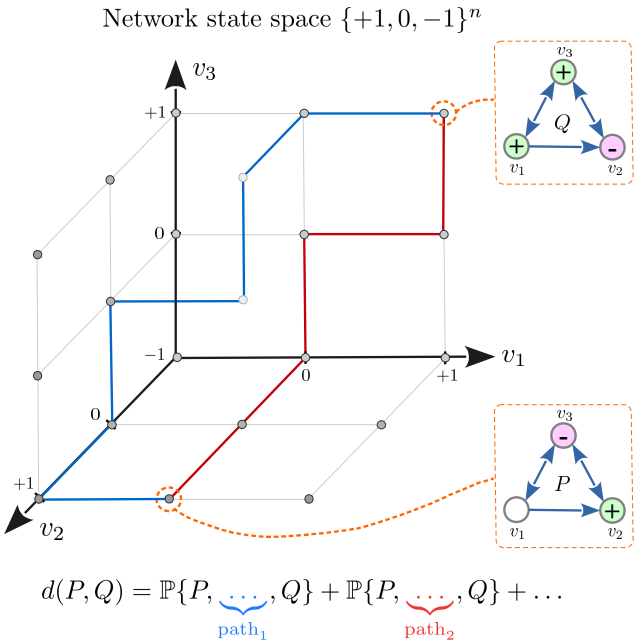


Fig. 1. $d_{\mathcal{M}}(P,Q)$ accumulates likelihoods of all possible network state evolution paths.

Despite the attractive semantics of $d_{\mathcal{M}}$, its computation is clearly unfeasible, as the number of possible evolution paths between two network states is exponential in the number of nodes. To come up with a tractable alternative, we simplify $d_{\mathcal{M}}$ by making several assumptions.

Assumption 1 (Maximum-likelihood opinion evolution). *Opinions evolve along to the most likely network state evolution path.*

According to Assumption 1—standard for maximum likelihood estimation methods—we will not care about every possible network state evolution path, instead, focusing on the most likely one. Under this Assumption $d_{\mathcal{M}}$ simplifies to $d^{(1)}$ as follows:

$$d^{(1)}(P,Q) = \max_{k\geq 2} \max_{\substack{X(1,\ldots,k)\\ X(1)=P, X(k)=Q}} \mathbb{P}\{X(1,\ldots,k)\}. \tag{4}$$

To further simplify the obtained distance measure, we will target the term being maximized in (4), and make another assumption.

Assumption 2 (Independent Markovian opinion acquisition). *Users acquire opinions asynchronously, independently of other users' opinion acquisition, relying only on the opinions of network users at the previous time.*

According to this assumption, opinions of different users evolve at individual time scales (in contrast to synchronous opinion evolution, where all users simultaneously update their opinions), and depend only on the previously observed opinions of other users. This assumption simplifies distance measure $d^{(1)}$ into $d^{(2)}$ as follows.

$$d^{(2)}(P,Q) = \max_{k\geq 2} \max_{\substack{X(1,\ldots,k)\\ X(1)=P, X(k)=Q}} \prod_{i=1}^{n}\prod_{t=2}^{k} \mathbb{P}\left\{X_i(t) \mid X(t-1)\right\} \tag{5}$$

Now, to make the maximization task in (5) tractable, we will make one more assumption.

Assumption 3 (Opinion source uniqueness). *An opinion is adopted by a user from a single most likely source.*

This assumption—previously used by Gomez-Rodriguez et al. [21] in the context of cascade inference—is natural in those cases when obtaining knowledge immediately causes or is equivalent to opinion acquisition, such as in the case of learning an incriminating piece of news about a politician. In these situations, a contact with a single information source is sufficient to acquire opinion, and contacting additional sources would not solidify that opinion even further.

If we put $f_{ji} \in [0,1]$ to be the likelihood of user j being the source for opinion acquisition by user i, $d^{(2)}$ simplifies under Assumption 3 into $d^{(3)}$ as follows:

$$d^{(3)}(P,Q) = \prod_{v\in\{+1,-1\}} \max_{\{f_{ji}\}} \prod_{i\in I_Q^v} \sum_{j\in I_P^v} f_{ji}\mathbb{P}\left\{\begin{array}{l} j \text{ infects } i \text{ with opinion } v \text{ along the most likely}\\ \text{evolution path } P = X(1), X(2), \ldots, X(\cdot) = Q\end{array}\right\}, \tag{6}$$

$$\sum_j f_{ji} = |Q_i|,\ \sum_i f_{ji} \in \mathbb{Z}^+,$$

where I_P^v is the set of users holding opinion v in network state P. Here, f_{ji} can alternatively be viewed collectively as a probabilistic mapping between opinion sources and destinations. The obtained distance measure (6) is equivalent to the following one expressed using log-likelihoods

$$\begin{aligned} d^{(4)}(P,Q) &= \sum_{v\in\{+1,-1\}} \min_{\{f_{ji}\}} \sum_{i\in I_Q^v}\sum_{j\in I_P^v} f_{ji}\left(-\log\mathbb{P}\left\{\begin{array}{l} j \text{ infects } i \text{ with opinion } v \text{ along the most likely}\\ \text{evolution scenario } P = X(1), X(2), \ldots, X(\cdot) = Q\end{array}\right\}\right)\\ &= \sum_{v\in\{+1,-1\}} \min_{\{f_{ji}\}} \sum_{i\in I_Q^v}\sum_{j\in I_P^v} f_{ji}D_{ji}(P,v), \end{aligned} \tag{7}$$

$$\sum_j f_{ji} = |Q_i|,\ \sum_i f_{ji} \in \mathbb{Z}^+. \tag{8}$$

In the expression above, $D_{ji}(P, \upsilon) \in \mathbb{R}^{+n \times n}$ is the log-likelihood (or cost) of opinion υ's spreading from user j to user i along the most likely path through the network in state P. Provided that for each pair of nodes j and i, the chosen opinion dynamics model $\mathcal{M}$ defines the likelihood $\mathbb{P}_{ji}(P, \upsilon)$ of opinion υ spreading through edge (j, i) in network state P—as per (1)—$D_{ji}(P, \upsilon)$ is defined as *the length of the shortest path from node j to node i in the network whose structure is identical to that of the network that P is defined over, and whose edge (j, i) is weighted with $-\log \mathbb{P}_{ij}(P, \upsilon)$.*

Finally, if we relax $f_{ji} \in [0, 1]$ in (7) to be arbitrary non-negative reals, the obtained distance measure (7)-(8) almost exactly matches Earth Mover's Distance (EMD) described in detail in Sec. 2.3:

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

$$\{\widehat{f}_{ij}\} = \arg\min_{\{f_{ij}\}} \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}D_{ij},\ \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = \min\Big\{\sum_{i=1}^{n} P_i, \sum_{j=1}^{m} Q_j\Big\},$$

$$f_{ij} \geq 0,\ \sum_{j=1}^{m} f_{ij} \leq P_i,\ \sum_{i=1}^{n} f_{ij} \leq Q_j,\ (1 \leq i \leq n, 1 \leq j \leq m).$$

Besides the difference in the normalization factor $\sum_{i,j} \widehat{f}_{ij}$, EMD imposes an extra constraint upon the sum of f_{ij}, requiring $\sum f_{ij} = \min\{\sum_i P_i, \sum_j Q_j\}$. For us, it would roughly mean that the number of active users—holding opinions +1 and −1—in both network states P and Q should be equal. The latter, however, does not hold in practice—as shown in Fig. 2—as already active users may be unwilling to become neutral, and even more active users may appear while information spreads through the network. This difference will disappear when we replace EMD with its generalization EMD* in Sec. 4.

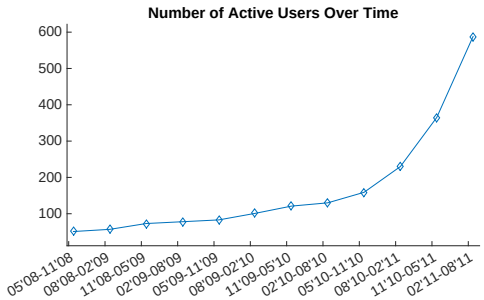


Fig. 2. Growth of the number of active users—holding non-neutral opinions w.r.t. topic "Obama"—in a sample of Twitter network over time.

Based on the obtained expression (7) and expression (2) for EMD, we can define the nonsymmetric version of our Social Network Distance as follows

$$\mathrm{SND}^{asym}(P, Q) = \sum_{\upsilon \in \{+1, -1\}} \mathrm{EMD}(P^\upsilon, Q^\upsilon, D(P, \upsilon)),$$

where P^υ is a network state containing absolute values of the entries of P, in which all the users holding opinions different from υ in P are considered neutral (Q^υ is defined similarly), and ground distance $D(P, \upsilon)$ consists of log-likelihoods as defined in (7). The defined above $\mathrm{SND}^{asym}(P, Q)$ is suitable for the comparison of time-ordered network states—when we know that network state P preceded network state Q in time. However, we would like to be able to compare arbitrary unordered

network states. This may be important for such applications as nearest neighbor network state search, where nearest neighbor candidate network states may either precede or succeed in time the target network state. To that end, instead of building upon $d_{\mathcal{M}}(P,Q)$, we will be interested in $\sqrt{d_{\mathcal{M}}(P,Q)\cdot d_{\mathcal{M}}(Q,P)}$, and, having repeated all the simplifications with the obtained expression, define our Social Distance Measure as follows.

Definition 3.1 (Social Network Distance (SND)).

$$\boxed{\mathrm{SND}(P,Q)=\frac{1}{2}\sum_{v\in\{+1,-1\}}\Big[\mathrm{EMD}(P^v,Q^v,D(P,v))+\mathrm{EMD}(Q^v,P^v,D(Q,v))\Big],} \tag{9}$$

where EMD is a version of Earth Mover's Distance. In Appendix A.5, we provide a toy example of computing SND for two network states, where in place of EMD we use its generalization EMD* designed in Sec. 4.

Notice that SND is a linear combination of multiple instances of EMD, so the following theorem trivially holds.

Theorem 3.2 (Metricity of SND). *SND is metric as long as the underlying EMD is metric.*

Also note that, since SND is defined via log-likelihoods, as per (7)-(9), its relationship with the likelihood of a network's transitioning between two network states is inverse—higher distance values correspond to lower likelihoods, and vice versa.

To summarize, we have defined Social Network Distance (SND)—a distance measure that approximates the (log-)likelihood of a network state's most-likely transition into another network state. As we have mentioned earlier, SND was defined via EMD ignoring the normalization factor in the definition (2) of EMD as well as the difference in the constraints between (2) and (8). In the following Section 4, we will generalize EMD to address both these issues, and use its generalization EMD* in definition (9) of SND.

4 GENERALIZED EARTH MOVER'S DISTANCE (EMD*)

4.1 Why do we need a new Earth Mover's Distance?

Our distance measure SND defined in the previous Sec. 3 uses an EMD, such as the original Earth Mover's Distance [46], as a building block. Unfortunately, the original EMD cannot adequately compare network states P and Q having different total masses, that is, network states with $\sum P_i \neq \sum Q_j$—it ignores the mass mismatch $|\sum P_i - \sum Q_j|$, so that a network state with a very small total mass has a very small distance to *any* other network state. Applicably to states of a social network, that limitation is particularly pronounced, as, usually, subsequent network states have more active users and, hence, a larger total mass than preceding network states.

There are several versions of EMD that address the original EMD's neglect for network state mass mismatch. One of them [41] augments EMD with an additive mass mismatch penalty as

$$\widehat{\mathrm{EMD}}(P,Q,D)=\mathrm{EMD}(P,Q,D)\min\left\{\sum_{i=1}^n P_i,\sum_{j=1}^n Q_j\right\}+\alpha\max_{i,j}\{D_{ij}\}\left|\sum_{i=1}^n P_i-\sum_{j=1}^n Q_j\right|,$$

where α is a constant parameter. Term $\alpha\max_{i,j}\{D_{ij}\}|\sum P_i-\sum Q_j|$ in the above expression represents the mass mismatch penalty that depends only on the magnitude of the mass mismatch and the maximum ground distance, thereby, being unable to capture the fine details of the network's structure that D can depend upon. This is, however, inadequate for the comparison of the states of a social network, because the network's behavior depends not only on the number of new activations, but as importantly on where these newly activated users are located in the network.

Another Earth Mover's Distance version EMD$^{\alpha}$ [34] extends each network state with an extra node—*the bank node*—whose value is chosen to equalize the network states' total masses. An example of such an extension is shown in Fig. 3.

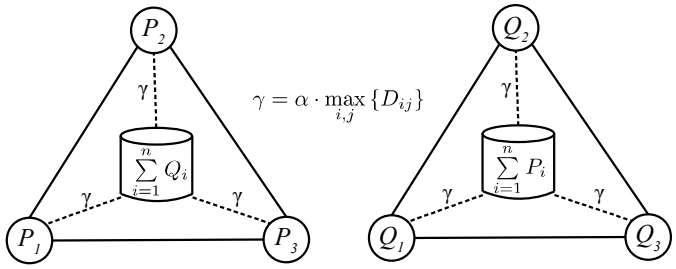


Fig. 3. Network states P and Q defined over the same network are extended with bank nodes, whose values are chosen, so that the total masses of the extended network states $\widetilde{P}$ and $\widetilde{Q}$ are equal. The ground distances $\widetilde{D}_{bank,i} = \widetilde{D}_{i,bank} = \gamma$ from and to the bank node are uniformly defined based on the largest ground distance between the initially present nodes.

However, as we establish in Theorem A.1 in the Appendix, EMD$^{\alpha}$ is numerically equivalent to $\widehat{\text{EMD}}$ and, hence, is also inadequate for the purpose of network state comparison for the same reason $\widehat{\text{EMD}}$ is.

In Sec. 6.4, we will show how the above statements about inadequacy of existing versions of EMD for opinion evolution analysis translate into performance of our anomaly detection method using different versions of EMD.

Hence, we need to design a new Earth Mover's Distance-like primitive that would fit the comparison of states of a social network, and replace EMD in SND's definition (9).

4.2 Generalized Earth Mover's Distance (EMD*)

In this section, we propose EMD*—a new version of Earth Mover's Distance, building upon EMD$^{\alpha}$'s idea of augmenting network states to even their masses. However, unlike its predecessor, EMD* extends network states with *multiple local bank nodes*—as shown in Fig. 4—and distributes the total mass mismatch over all of them, thereby, relating the mass mismatch penalty to the structure of the network, while achieving the total mass equality of the two network states under comparison.

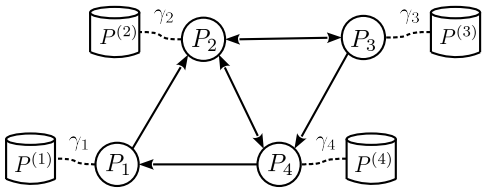


Fig. 4. A network state $[P_1, \ldots, P_4, P^{(1)}, \ldots, P^{(4)}]$ extended with *local bank nodes* $P^{(i)}$. The undirected edges to and from the bank nodes—displayed dashed—are weighted with ground distances $\gamma_1, \ldots, \gamma_4$, respectively.

Prior to formalizing EMD*, let us, first, better understand its advantage over the existing EMDs as well as fitness to the analysis of opinion evolution. Consider the example in Fig. 5. There are three network states defined over the same network, which has two pronounced clusters L and R connected by three bridge edges. The distribution of mass over cluster L is identical in all three

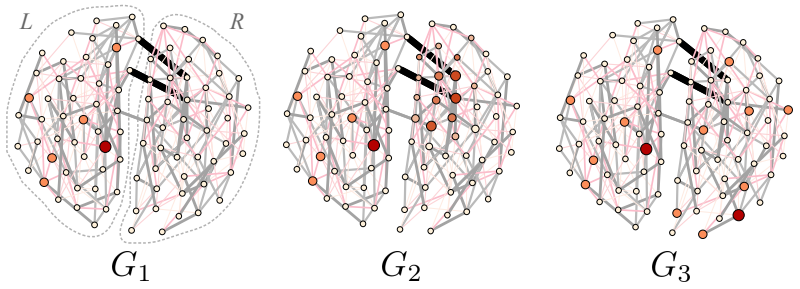


Fig. 5. Three network states defined over the same two-cluster network.

states G_i, while cluster R is empty in G_1 and has some differently distributed mass in G_2 and G_3. In G_2 the extra mass has been "propagated" from cluster L to cluster R through the bridges, while in G_3 the same amount of extra mass has been randomly distributed over cluster R. Thus, if we assume that G_2 and G_3 have "evolved" from G_1 through a (not completely random) network process of mass diffusion, then G_2 should intuitively be closer to G_1 than G_3 is. However, only EMD* captures this intuition as $\text{EMD}^\star(G_1, G_2) < \text{EMD}^\star(G_1, G_3)$, while for EMD$^\alpha$ and $\widehat{\text{EMD}}$, G_2 and G_3 are equidistant from G_1, and for the original EMD, both G_2 and G_3 are identical to G_1. In Sec. 6.4, we will show how this intuition translates into the relative performance of different versions of EMD when used with our method for anomaly detection.

We will now turn from the intuition for the formal definition of different components of EMD* and, eventually, EMD* itself. The extension of network state $P \in \mathbb{R}^n$ with local bank nodes requires definition of the ground distances γ_i to/from the bank nodes as well as the banks' values $P^{(i)}$.

▷ *Bank node ground distances γ_i:* In the extreme case, when $\gamma_i = 0$, in the transportation problem underlying EMD*, the mass is transported to/from the banks at a zero cost, which would result in EMD*'s ignorance of the network state mass mismatch Δ, making it similar to the original EMD up to normalization by Δ. If, on the other hand, γ_i is much larger than the ground distances between regular (non-bank) nodes, then the value of EMD* will be dominated by the effect of the network state mass mismatch, hiding the impact of the actually present mass. Thus, γ_i should be chosen of the same order as the ground distances D_{ij} between regular nodes, with the particular values of γ_i being empirically learned.

▷ *Bank node values $P^{(i)}$:* The values of the added bank nodes should be determined based upon two ideas. Firstly, the value of a bank node should intuitively be proportional to the total mass of the node the bank is attached to, thereby, preserving the relative distribution of mass over the network. Secondly, the values of all the bank nodes should be such, that the two network states under comparison have equal total masses. The following definition of value $P^{(i)}$ of a bank node connected to the i'th node of network state P in the context of comparing network states $P = [P_1, \dots, P_n]$ and $Q = [Q_1, \dots, Q_n]$ incorporates both above mentioned requirements.

$$P^{(i)} = \begin{cases} \left(\sum_{j=1}^n Q_j / \sum_{j=1}^n P_j - 1\right) P_i, & \text{if } \sum Q_j > \sum P_j, \\ 0, & \text{otherwise.} \end{cases}$$

Next, we formally define EMD*. Suppose we are given two network states $P = [P_1, \dots, P_n]$ and $Q = [Q_1, \dots, Q_n]$ defined over a network $G = \langle V, E \rangle$ with ground distance $D_{n \times n}$. Network states P and Q are extended with bank nodes, with $P^{(i)}$ and $Q^{(i)}$ being the values of the bank nodes attached to the i'th regular node of P and Q, respectively. Ground distances to/from the bank nodes are defined, collectively, as $\gamma = [\gamma_1, \dots, \gamma_n]^\intercal$.

Then, EMD* is defined as follows.

Definition 4.1 (Generalized Earth Mover's Distance (EMD*)).

$$\text{EMD}^\star(P,Q) = \text{EMD}(\widetilde{P},\widetilde{Q},\widetilde{D}) \max\left\{\sum P_i, \sum Q_j\right\}, \tag{10}$$

$$\widetilde{P} = \left[P, P^{(1)}, \ldots, P^{(n)}\right], \quad \widetilde{Q} = \left[Q, Q^{(1)}, \ldots, Q^{(n)}\right],$$

$$\widetilde{D} = \left[\begin{array}{c|c} D & D + \mathbb{1}_n \otimes \gamma^\intercal \\ \hline D + \mathbb{1}_n^\intercal \otimes \gamma & D + \mathbb{1}_n \otimes \gamma^\intercal + \mathbb{1}_n^\intercal \otimes \gamma - 2\,\text{diag}(\gamma) \end{array}\right],$$

where $P^{(i)}$ is the value of the i'th bank node that P is extended with (same for Q), $\mathbb{1}_n \in \mathbb{R}^{n\times 1}$ is a vector of all ones, $\text{diag}(v)$ is a diagonal matrix with the elements of vector v on its main diagonal, and $\otimes$ is Kronecker product.

Metricity of EMD*, which can be exploited to improve practical performance of distance-based search in applications [13], is established in the following Theorem, proven in Appendix A.3.

Theorem 4.2. *Given a finite set $\mathcal{H}$ of network states and metric ground distance D, EMD* defined over D is metric on $\mathcal{H}\times\mathcal{H}$.*

Having generalized EMD*, so that it can handle comparison of network states, we will now be using EMD* as our Earth Mover's Distance of choice in the definition (9) of Social Network Distance. We provide a toy example of computing EMD* as part of SND in Appendix A.5.

5 EFFICIENT COMPUTATION OF SND

While we have designed SND (9), its computation is non-trivial. Since SND is a linear combination of multiple instances of EMD*, its computation involves:

- Computing the ground distance $D(G(t), v)$ based on the structure of the underlying network $G = \langle V, E\rangle$ ($|V| = n$, $|E| = m$) and the opinions of the users in network state $G(t)$.
- Computing EMD*, when the network states and the ground distance are provided.

Computing the ground distance D implies computing shortest paths in a network with edge weights $-\log \mathbb{P}_{ij}(P, v)$. Direct all-to-all shortest path computation using Dijkstra's algorithm would incur time cost $O(n^2 \log n)$ for sparse G. Computing EMD* is algorithmically equivalent to computing EMD, and, since the latter is formulated as a solution to a transportation problem, it can be computed either using a general-purpose linear solver, such as Karmarkar's algorithm [27], or a solver that exploits the special structure of the transportation problem, such as the transportation simplex algorithm [25]. The time complexity of both these algorithms, however, is supercubic in n. Thus, the exact computation of SND using existing techniques is prohibitively expensive at the scale of real-world online social networks. Furthermore, the existing approximations of EMD are either inapplicable to the comparison of network states derived from a social network's states [33, 41, 42], since they drastically simplify the ground distance [31, 50], or are effective only for some graphs, such as trees, structurally not characteristic of social networks [37].

Nevertheless, in what follows, we propose a method to compute SND exactly and in time linear in n under the following two realistic *assumptions*.

Assumption 1: The number n_Δ of users who change their opinions between two network states under comparison is significantly smaller than the total number n of users in the network.

Assumption 2: The log-likelihoods $-\log \mathbb{P}_{ij}(P, v)$ of opinion spread—being the edge weights in the network in which ground distances are defined as lengths of shortest paths—are positive integers bounded from above by constant $U \ll +\infty \in \mathbb{Z}^+$.

Assumption 1 is reasonable in most applications the network states under comparison are not very far apart in time and, hence, $n_\Delta \ll n$; Assumption 2 is reasonable, because most of the log-likelihoods are large reals, and rounding them would not introduce a large error.

We will now use the above stated assumptions to design an efficient algorithm for SND. Since, according to its definition (9), SND's computation involves computation of four instances of EMD*, we will actually be dealing with efficient computation of EMD* on the inputs supplied by SND. Our method for efficient computation of SND requires the following two lemmas.

Lemma 5.1. *For any two network states $P \in \mathbb{R}^n$ and $Q \in \mathbb{R}^n$, and ground distance $D \in \mathbb{R}^{n\times n}$, if $P_i = Q_i = 0$, then the removal of i'th elements from P, Q, as well as i'th row and column from D does not affect the value of EMD$^\star(P, Q, D)$.*

Lemma 5.1 is straightforward, since zero-value P_i and Q_i do not supply or consume any mass in the underlying transportation problem, and, hence, do not affect the cost of the optimal transportation plan. While Lemma 5.1 allows removing redundant suppliers and consumers from the underlying transportation problem, the following Lemma 5.2—proven in Appendix A.4—allows to transform the network states, without affecting the value of EMD*, exposing the redundant suppliers and consumers for removal.

Lemma 5.2 (Network State Reduction). *Given two arbitrary network states $P, Q \in \mathbb{R}^n$ and a ground distance $D \in \mathbb{R}^{n\times n}$, if D is semimetric*[2]*, then for any $i \in \{1, \dots, n\}$,*

$$EMD^\star(P, Q, D) = EMD^\star([P_1, \dots, P_{i-1}, P_i - \min\{P_i, Q_i\}, P_{i+1}, \dots, P_n], \\ [Q_1, \dots, Q_{i-1}, Q_i - \min\{P_i, Q_i\}, Q_{i+1}, \dots, Q_n], D).$$

We will, now, state the main result for the efficient computation of SND as Theorem 5.3, whose constructive proof provides the algorithm for SND's computation.

Theorem 5.3. *Under Assumptions 1 and 2, SND between network states $P = [P_1, \dots, P_n]$ and $Q = [Q_1, \dots, Q_n]$ defined over network $G = \langle V, E\rangle$, ($|V| = n, |E| = m$) can be computed in time*

$$T = O(n_\Delta(m + n\sqrt{\log U} + n_\Delta^2 \log(n_\Delta n U))).$$

Proof. Throughout this proof, we will use notation $P^+ = P^{(+1)}$ to denote a network state where users holding negative opinions are considered neutral, and $D(P, +) = D(P, +1)$ to denote the ground distance for the spread of opinion +1 through the network in state P, as well as the similar notation P^- and $D(P, -)$ for negative opinions.

We will focus on the efficient computation of the first summand EMD$^\star(P^+, Q^+, D(P, +))$ in definition (9) of SND(P, Q, D), as computation of three other summands is algorithmically equivalent and takes the same time. For the analysis of the computation of EMD$^\star(P^+, Q^+, D(P, +))$, let us assume, without loss of generality, that $\sum_{i=1}^n P_i^+ \geq \sum_{j=1}^n Q_j^+$. As per (10), EMD$^\star(P^+, Q^+, D(P, +))$ is the solution of a transportation problem with suppliers $\widetilde{P}^+ = [P_1^+, \dots, P_n^+, \mathbb{0}_{1\times n}]$, consumers $\widetilde{Q}^+ = [Q_1^+, \dots, Q_n^+, Q^{+(1)}, \dots, Q^{+(n)}]$, and ground distance $\widetilde{D}(P, +)$.

Now, we can apply Lemmas 5.1 and 5.2 to reduce the size of the obtained transportation problem. From Assumption 2, $\widetilde{D}(P, +)$ is semimetric. Non-negativity and identity of indiscernibles straightforwardly follow from Assumption 2 and the definition of the length of a shortest path. Subadditivity follows from the shortest path problem's optimal substructure. Thus, we can apply Lemma 5.2 to each pair $\widetilde{P}_i^+, \widetilde{Q}_i^+$ of corresponding suppliers and consumers, and due to Assumption 1, a large number $(n - n_\Delta)$ of them have equal values. As a result, many suppliers and consumers

[2] A semimetric is a metric with the symmetry requirement dropped

become empty. Then, due to Lemma 5.1, all the obtained empty nodes can be disregarded. If we put $M_i = \min\{P^+_i, Q^+_i\}$, then the reduced transportation problem is defined for suppliers $[P^+_{i_1} - M_{i_1}, \ldots, P^+_{i_{n_\Delta}} - M_{i_{n_\Delta}}]$ and consumers $[Q^+_{j_1} - M_{j_1}, \ldots, Q^+_{j_{n_\Delta}} - M_{j_{n_\Delta}}, Q^{+(1)}, \ldots, Q^{+(n)}]$, and ground distance $\widetilde{D}(P,+)$ that contains only the rows and columns corresponding to the remaining suppliers and consumers. The remaining suppliers and non-bank consumers correspond to the users who have different opinion in P^+ and Q^+, and the number of such users, due to Assumption 1, is at most n_Δ. The bank nodes $Q^{+(1)}, \ldots, Q^{+(n)}$, however, do not get affected by Lemma 5.2 in $\widetilde{Q}^+$ (since only the banks of the lighter network state P can have non-zero mass) and hence are not removed, yet, they are removed from $\widetilde{P}^+$ due to Lemma 5.1. Thus, we have an unbalanced transportation problem, where the number n_Δ of suppliers is much less than the number $n + n_\Delta$ of consumers.

Now, in order to compute $\text{EMD}^\star(P^+, Q^+, D(P,+))$, we need to compute $\widetilde{D}(P,+)$ and to actually solve the obtained transportation problem.

Due to the structure of the reduced transportation problem, we need to compute only a small part of $\widetilde{D}(P,+)$. Since there are at most n_Δ suppliers, we need to solve at most n_Δ instances of single-source shortest path problem (SSSP) with at most $n_\Delta + n$ destinations. Since, due to Assumption 2, edge costs in the network are integer and bounded by U, each SSSP instance can be solved using Dijkstra's algorithm based on a combination of a radix and a Fibonacci heaps [2] in time

$$T_{sssp} = O(m + n \log \sqrt{U}).$$

(Notice, that if we assumed $\sum_{i=1}^{n} P^+_i \leq \sum_{j=1}^{n} Q^+_j$, and the reduced $\widetilde{P}^+$ contained $n_\Delta + n$ nodes, we would *not* need to run $n_\Delta + n$ SSSP instances. Instead, we would invert the edges in the network and compute the shortest paths in reverse, still solving only n_Δ SSSP instances.)

Next, we approach the solution of the reduced transportation problem with known ground distances. This problem can be viewed as a minimum-cost network flow problem in an unbalanced bipartite graph, where the number of consumers is much greater than the number of suppliers or vice versa. Since, due to Assumption 2, edge costs are integers bounded by U, our minimum-cost flow problem can be solved using Goldberg-Tarjan's algorithm [20] augmented with the two-edge push rule of Ahuja et al. [3] in time

$$T_{transp} = O(n_\Delta m + n^3_\Delta \log (n_\Delta \max_{i,j} \widetilde{D}(P,+)_{ij})).$$

Since no shortest path has more than $(n-1)$ edge, and the edge costs are bounded by U, the expression for time simplifies to

$$T_{transp} = O(n_\Delta m + n^3_\Delta \log (n_\Delta n U)).$$

Thus, the total time for computing $\text{EMD}(P^+, Q^+, D(P,+))$ and, consequently, $\text{SND}(P, Q, D)$ is

$$T = O(n_\Delta T_{sssp} + T_{transp}) = O(n_\Delta (m + n \log \sqrt{U} + n^2_\Delta \log (n_\Delta n U))).$$

□

Theorem 5.3 immediately entails the following corollary.

Corollary 5.4. *If the social network is sparse, that is $m = O(n)$, and the number n_Δ of users who changed their opinions is bounded, then SND is computable in time $O(n)$.*

6 EXPERIMENTAL RESULTS

In this section, we report experimental results, demonstrating the utility of SND in applications in comparison to other distance measures. We also evaluate scalability of our implementation of SND.

6.1 Experimental Setup

Twitter Data: Our Twitter dataset is based on the crawled data of [35], and includes 48M tweets sent over 6 years. From these tweets, we select around 270k tweets sent between May-2008 and August-2011, containing hashtags related to the political topics—such as *"Obama"*, *"GOP"*, *"Palin"*, *"Romney"*—popular in the US during these years, and connect users in a network based on their follower-followee relationship. As a result, we obtain a network of 10k users tweeting about politics, each having an average of 130 neighbors in the network. Within each quarter, we quantify the sentiment of each tweet as described in [36] using the sentiment classification approach of [51]. Then, we find the users who posted at least as many as 10% of the average number of per-user tweets posted within the quarter, and label them as active. Other users are assumed to be neutral, and their opinions for the quarter are set to 0. An active user's opinion is set to $+1$ (-1) if he or she has posted at least 4 times more positive (negative) than negative (positive) tweets within the quarter, assuming that such a skew in the tweets' sentiment is enough to identify whether the user likes or dislikes the topic. Otherwise—if an active user has posted enough of both positive and negative tweets—this user's opinion is set to 0, that is, such user is considered neutral. As soon as we have quantified the quarterly opinion of each user, the opinions of all the users comprise that quarter's network state.

Synthetic Data: We also perform experiments on synthetic scale-free networks of sizes $|V|$ from 10k to 200k and scale-free exponents from -2.9 to -2.1. To generate the first network state, a number of initial adopters are chosen uniformly at random, and approximately equal numbers of them adopt opinions $+1$ and -1. Each subsequent network state $G(t+1)$ is randomly generated from the preceding network state $G(t)$ as follows. A number of $G(t)$'s neutral users get a chance to be activated. Each of them adopts an opinion from her neighbors with probability $\mathbb{P}_{nbr}$ and a random opinion with a smaller probability $\mathbb{P}_{ext}$. If a user is to adopt an opinion from the neighbors, which opinion to adopt is decided in a probabilistic voting fashion based on the numbers of active in-neighbors of each kind. This generative model is a version of Independent Cascade model [11], where edges in a neighborhood are activated simultaneously with probability $\mathbb{P}_{nbr}$, and external influence $\mathbb{P}_{ext}$ is allowed.

Distance Measures: In our experiments, SND does not make any assumptions regarding the above described generative process of opinion evolution in synthetic data, and assumes the simple model (1), whose parameter $c_{adverse}$, $c_{neutral}$, and $c_{friendly}$ values are learned from how well SND performs in applications, and, as a result, are set, respectively, to 1000, 40, and 5 for anomaly detection experiments, and to 100, 20, and 5 in opinion prediction experiments. SND is compared with the following distance measures.

- $hamming(P,Q)$. Hamming distance is a representative of ℓ_p-like distance measures performing basic coordinate-wise comparison. It measures the number of users whose opinions differ in network states P and Q.
- $quad\text{-}form(P,Q,L) = \sqrt{(P-Q)^\intercal L(P-Q)}$. Quadratic-Form Distance [23] based on the Laplacian matrix L [39] of the network. It combines the differences of opinions of the corresponding users based on the network's structure. More specifically, when $L = \mathrm{diag}(A\mathbb{1}) - A$ is the difference between the diagonal degree matrix $\mathrm{diag}(A\mathbb{1})$ of the network and its adjacency matrix A, $\sqrt{(P-Q)^\intercal L(P-Q)} = \sum_{(i,j)\in E} A_{ij}((P-Q)_i - (P-Q)_j)^2$ aggregates differences between P and Q along all the edges present in the network.
- $walk\text{-}dist(P,Q) = \frac{1}{n}\|con(P) - con(Q)\|_1$. Compares vectors $con(P) = [con(P_1), \ldots, con(P_n)]$ of users' "contention", where $con(P_i)$ is the amount by which the i'th user's opinion deviates

from the opinion of this user's average active in-neighbor. Thus, *walk-dist* summarizes how different the network's users are from their respective neighbors.

6.2 Detecting Anomalous Network States

Synthetic Data: In a series $G(1), \ldots, G(t), \ldots$ of network states, we want to detect which transitions in the series are anomalous, that is, when opinions change unexpectedly deviating from their established evolution pattern. In particular, we are interested in those anomalies that are hard to detect by observing simple summaries of social network states, such as the number of newly activated users. To simulate such anomalies with synthetic data, we change the values of $\mathbb{P}_{nbr}$ and $\mathbb{P}_{ext}$—controlling the process of opinion evolution from one network state to the other—preserving their sum, thereby, affecting which users get activated, yet, maintaining the same activation rate.

To detect anomalies, in a series of network states, we compute the distances between adjacent states, normalize these distances by the number of active users, and rescale the obtained values to fit range $[0, 1]$. Then, spikes in the resulting series of distances are considered anomalies.

A qualitative analysis of anomaly detection on synthetic data is presented in Fig. 6. For each

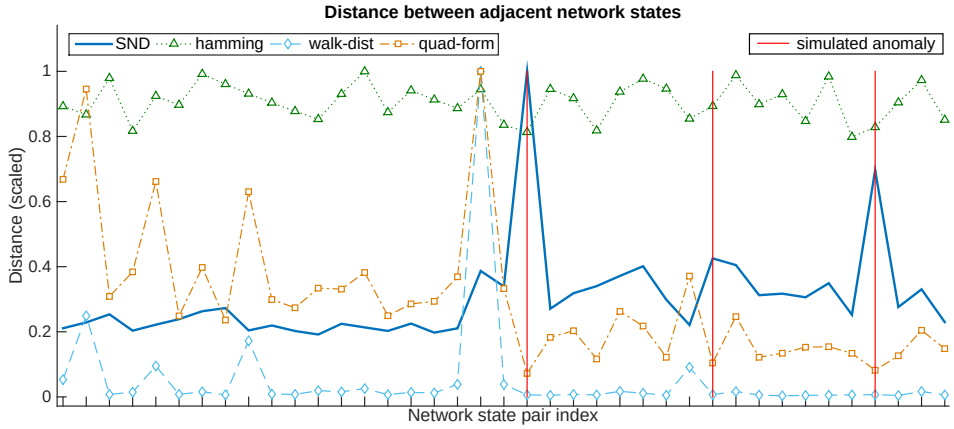


Fig. 6. Anomaly detection on synthetic data. $|V| = 20$k, scale-free exponent $\gamma = -2.3$. A series of 40 network states is generated using $\mathbb{P}_{nbr} = 0.12$ and $\mathbb{P}_{ext} = 0.01$ for normal and $\mathbb{P}_{nbr} = 0.08$ and $\mathbb{P}_{ext} = 0.05$ for anomalous network states' generation, respectively. The three simulated anomalies are displayed as solid vertical lines.

simulated anomaly, SND produces a well noticeable spike, reacting to the qualitative change in the opinion dynamics process, while other distance measures do not recognize such anomalies. The additional experiment exposing this difference in sensitivity of SND vs. simpler distance measures is provided in Sec. 6.5.

In order to quantify the performance of the competing distance measures at detecting simulated anomalies, we create a simple anomaly score $S_t = |(d_t - d_{t-1}) + (d_t - d_{t+1})|$, where d_t is the value of a given distance measure at time t normalized by the number of users active at time t and rescaled. The semantics of the above defined S_t—matching up to a constant factor the central finite-difference approximation of the second-order time derivative [30] of d_t—is such that large "hikes" in a distance series, e.g., $d_{t+1} \gg d_t$—large increases or decreases—receive high scores, and "spikes", e.g., $d_t \gg d_{t-1}$, $d_t \gg d_{t+1}$—large increases followed by large decreases or vice versa—receive even higher scores, making it easy to distinguish these two anomalous distance sequences from those close to linear. We rank the network state transitions for each compared distance

measure by S_t in decreasing order and compute true and false positive predictions for increasing ranks. We perform this evaluation on a series of 300 network states over a scale-free network with scale-free exponent $\gamma = -2.3$. 30 uniformly randomly selected pairs of adjacent network states are anomalous ($\mathbb{P}_{nbr} = 0.07$, $\mathbb{P}_{ext} = 0.011$), while the rest are normal ($\mathbb{P}_{nbr} = 0.08$, $\mathbb{P}_{ext} = 0.001$). The corresponding ROC and Precision-Recall curves are displayed in Fig. 7. SND's accuracy dominates

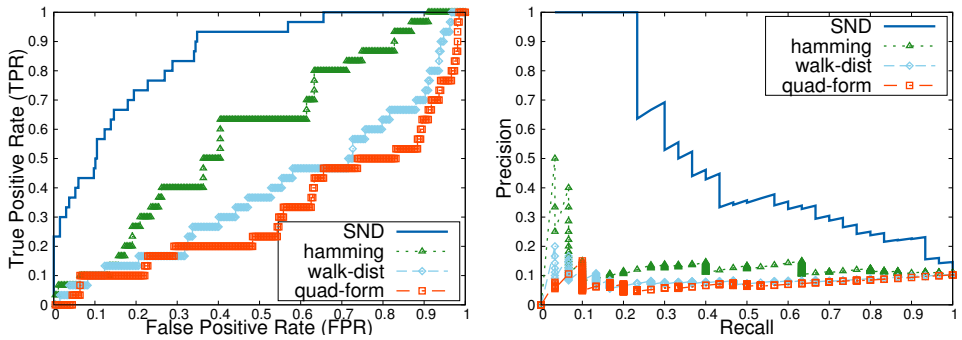


Fig. 7. ROC and Precision-Recall curves comparing the quality of anomaly detection by different distance measures in a series of 300 network states—in which 30 transitions are anomalous, while the rest are normal—over synthetic network with $|V| = 30k$ and scale-free exponent $\gamma = -2.3$. The network states are generated using $\mathbb{P}_{nbr} = 0.08$ and $\mathbb{P}_{ext} = 0.001$ for normal and $\mathbb{P}_{nbr} = 0.07$ and $\mathbb{P}_{ext} = 0.011$ for anomalous instances.

that of competing distance measures throughout the spectrum of false positive rates. Particularly, for false positive rates up to 0.3, SND achieves a true positive rate of 0.83, while the next best distance measure (*hamming*) achieves only 0.4. The same predictions presented in precision-recall space similarly demonstrate SND's dominance. It retains perfect precision for recall up to 0.2. Its precision then decreases while maintaining more than 2-fold improvement over alternatives for recall up to 0.95.

Twitter Data: To obtain the ground truth for anomaly detection on our Twitter dataset, we collect "search interest" data from Google Trends[3] and cross-check this data with American Presidents[4] log of political events in the US. The anomaly detection results for topic "Obama" are shown in Fig. 8.

We can distinguish two types of events based on SND's behavior relatively to that of other distance measures. One type is the polarizing events when SND noticeably disagrees with the other distance measures. For example, during quarters 05'09-11'09, the Economic Stimulus Bill had a highly polarized response in the House of Representatives[5], with no Republican voting in its favor. Another such anomaly takes place during quarters 02'10-08'10, when the Affordable Care Act ("Obama Care") was introduced, and which was a very controversial topic based on the House vote distribution[6] and the analysis from socialmention.com[7].

The other events are those where SND agrees with the other distance measures. Three examples are (a) "election", (b) "Tax plan", and (c) "bin Laden" (even though, all distance measures noticeably increase their value during the last quarter, we do not mark this quarter as anomalous, since we do not have the distance values for the next quarter.)

The (a) election of Barack Obama as the President of the US, extensively covered by the news media, had likely been accompanied by a very noticeable change in the rate of new user activation,

[3] http://www.google.com/trends/explore
[4] http://www.american-presidents-history.com
[5] http://www.nytimes.com/2009/01/29/us/politics/29obama.html
[6] https://tinyurl.com/obamacare-house-vote
[7] http://socialmention.com/search?q=obama+care

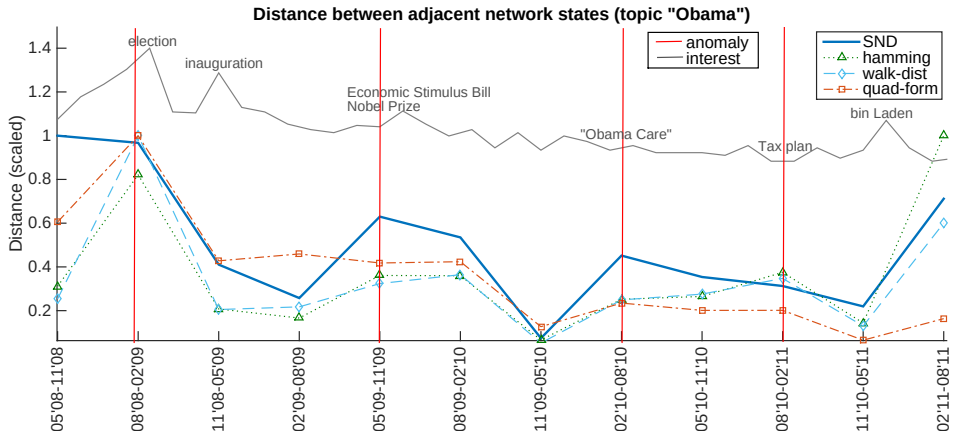


Fig. 8. Anomaly detection on Twitter data (May'08-Aug'11). The distance series are accompanied by the curve showing Google Trends' scaled interest in topic "Obama". Network states detected to be anomalous by at least one distance measure are displayed as solid vertical lines.

so, as expected, both SND and simpler distance measures sensitive to the user activation rate successfully detect this anomaly. However, the (b) Obama's tax cut extension and (c) bin Laden's assassination—not flagged as anomalies by SND—were not polarizing, as the tax cut had received large support in the Senate from both Democrats and Republicans[8], while bin Laden's assassination has probably evoked the same type of sentiment on Twitter across the US.

6.3 Predicting User Opinions

Our distance measure-based method for user opinion prediction—illustrated in Fig. 9—is as follows.

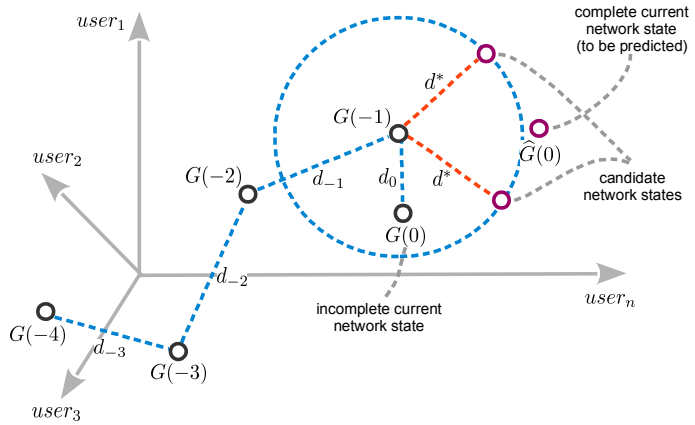


Fig. 9. Distance measure-based user opinion prediction. Network states $G(t)$ reside in the network state space. The series of distances $\langle \ldots, d_{-3}, d_{-2}, d_{-1} \rangle$ between past network states $\ldots, G(-4), G(-3), G(-2), G(-1)$ adjacent in time is extrapolated to estimate the distance d^* to the true current network state $\widehat{G}(0)$ to be predicted. We, then, search for the assignment of opinions to target users in $G(0)$ to make the distance from $G(-1)$ to the obtained network state $G^*(0)$ as close to d^* as possible.

[8] http://tinyurl.com/wiki-obama-tax-relief-2010

Given a series of states $\langle \ldots, d_{-3}, d_{-2}, d_{-1} \rangle$ of a social network, we want to predict the unknown opinions of a specified set of users in the true current network state $\widehat{G}(0)$ based on the observed recent $G(-t)$ $t = 1, 2, \ldots$ and the incomplete (partially known) current $G(0)$ network states. Such situation may arise either if the target users keep their profiles private or simply have not yet generated enough content in the current quarter to reliably quantify their opinions. We assume that over the recent time period—corresponding to the observed network states—the opinions in the network evolved "smoothly", so that the observed network states carry enough information to complete the partially known current network state. Under this assumption, and having chosen a distance measure $dist$, we compute the distances $d_{-t} = dist(G(-t-1), G(-t))$ between adjacent past network states, then, extrapolate the obtained series of distances via linear 4-point[9] least squares fitting to estimate the expected distance d^* from the most recent $G(-1)$ to the yet not fully known true current network state $\widehat{G}(0)$. Then, we search for the assignment of opinions to the target users in the partially known current network state $G(0)$—resulting in a candidate network state $G^*(0)$—that would make the distance $dist(G(-1), G^*(0))$ from the most recent to the candidate network state as close to the estimate d^* as possible.

While there may be multiple candidate network states $G^*(0)$ whose distances are close to d^*—the network states in Fig. 9 close to the sphere of radius d^* centered at $G(-1)$—due to the spatially-sensitive nature of network state comparison that SND provides, the set of such candidate network states will be rather small, and the opinions of the target users in these network states will be close to the true target user opinions in $\widehat{G}(0)$.

We have used two methods for the search of the best opinion assignment. The first method is a randomized search with uniformly randomly chosen opinions, and the number of random opinion assignments (100 in our experiments) being considerably lower than the total number of possible assignments (1M+ in our experiments). The second method was greedy hill climbing, the results for which are not reported, performing no better than the randomized search.

In each experiment, we uniformly randomly select 20 active users—with roughly equal representation of positive and negative opinions—in the current network state, predict their opinions and measure the prediction accuracy. We repeat this procedure 10 times, each time targeting a different set of users, and report means and standard deviations of the obtained prediction accuracies.

The predictions are made using the above distance measure-based method with SND as well as other distance measures. To put the prediction performance of these methods in context, we include in the comparison several non-distance measure-based opinion prediction methods.

- *icc-simulation*, *ltc-simulation* [40] simulate the model—Independent Cascade or Linear Threshold with uniformly randomly chosen thresholds, respectively—until convergence multiple times (from 10 to 500, in our experiments) and use the modes of the target users' opinion as the prediction. In our experiments, the simulation starts with the most recent completely known network state G_{-1}, and proceeds until 99.99% of users get active. The edge activation probabilities $\mathbb{P}_{edge}$ are selected uniformly, with $\mathbb{P}_{edge}$ ranging from 0.001 to 0.01. The results for the best $\mathbb{P}_{edge}$ are reported.
- *icc-max-likelihood*, *ltc-max-likelihood* are max-likelihood-based methods similar to [47] and [17]. Like SND-based opinion prediction, this method generates uniformly random opinion assignments, computes the likelihood of each resulting network state and uses the most likely one for the prediction. The opinion adoption likelihoods are computed as described in Appendix A.1, with all edges assumed to be active, and $\varepsilon = 0.01$.

[9] We have experimented with different numbers of points used with least squares. Using less than 4 points resulted in a poor opinion prediction performance, while using more than 4 points did not improve the performance much.

- *community-lp* [15, IV.B] detects communities in the network via label propagation and, then, predicts user opinions based on these users' membership in the discovered communities. This method does not rely on any opinion dynamics model, and only assumes that users likely connect with other likeminded users.

We experiment with both synthetic and Twitter data. For synthetic data, we generate a scale-free network with $n = 10,000$ users and scale-free exponent $\gamma = -2.5$. A series of network states is generated using the same version of Independent Cascade model as was used in anomaly detection experiments, with probabilities of opinion adoption from the neighbors $\mathbb{P}_{nbr}$ and from the external source $\mathbb{P}_{ext}$ ranging between 0.001 and 0.2. The number of initially active users is set to 800.

The opinion prediction results are summarized in Table 2. There are four important observations:

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 [40]	**76.25**	**9.54**	59.38	4.17
ltc-simulation [40]	67.50	11.65	58.75	5.18
icc-max-likelihood [47]	67.41	7.03	57.50	8.02
ltc-max-likelihood [47]	57.50	8.45	55.63	11.78
community-lp [15]	65.25	9.43	56.87	8.43

Table 2. Means μ and standard deviations σ of user opinion prediction accuracies.

(a) Among the distance-based methods, SND always performs best on both synthetic and Twitter data, with the mean prediction accuracy of 74-75% and a consistently low standard deviation. This suggests that SND captures more opinion dynamics-specific information than other distance measures, and should be preferred, particularly, when such simple statistics as the rate of new user activation are uninformative.

(b) Among the non-distance measure-based methods, *icc-simulation*'s prediction accuracy on synthetic data is 76.25%, the best result, comparable to 74.33% accuracy of SND. Such good performance of *icc-simulation* on synthetic data is not surprising, as its predictions rely on the Independent Cascade model, whose version was used to generate the synthetic data. On Twitter data, however, *icc-simulation*'s accuracy is 59.38%, compared to 75.63% accuracy of SND.

(c) Methods *icc-max-likelihood* and *ltc-max-likelihood* perform worse than SND, as they base prediction on the opinions of the closest active neighbors, while SND is looking for the most likely opinion propagation through potentially long paths in the network.

(d) SND-based method outperforms *community-lp* based on community detection via label propagation. In our experiments, *community-lp*'s prediction accuracy is 57-65%, while this method's authors report the accuracy of 95% for their data [15]. The likely cause of such a discrepancy is a *very* high level of homophily in their data (the reasons of which were discussed in [14]), while in our less homophilous data, *community-lp* performs worse by capturing only users' reachability by the opinions of each kind, whereas SND performs better by looking for the *most likely* opinion propagation scenario.

6.4 Anomaly Detection and Social Network Distance via EMD*, EMD$^\alpha$, and EMD

In Sec. 4.2, we motivated the design of a new version of Earth Mover's Distance by providing intuition for why its existing variants would not suit the analysis of opinion evolution. Here, we show how different versions of EMD perform in anomaly detection on synthetic data to provide additional experimental support to the claim that EMD* is a necessary component of SND. To that end, we borrow the experimental design from Sec. 6.2, and compare how SND—that internally uses EMD*—performs in anomaly detection on synthetic data in comparison to other implementations of SND that rely on EMD and EMD$^\alpha$ (and, hence, $\widehat{\text{EMD}}$, which is equivalent to EMD$^\alpha$), respectively. As computation of EMD and EMD$^\alpha$ does not scale, we perform this experiment on a network with 256 nodes. The results are show in Fig. 10.

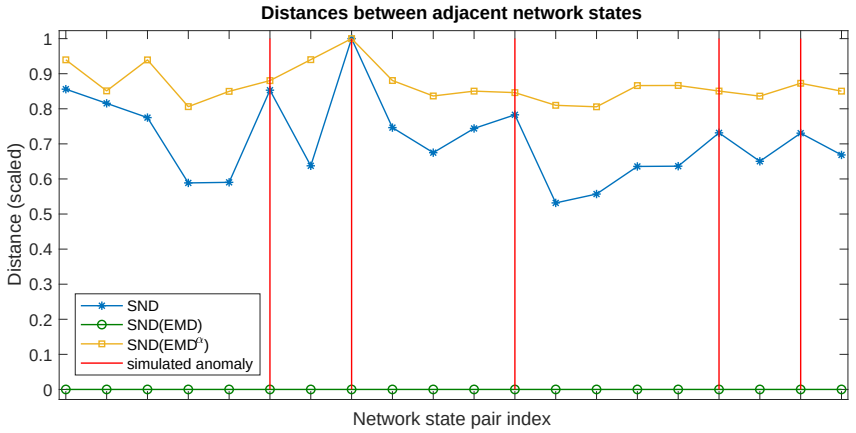


Fig. 10. Comparison of several versions of SND internally using EMD*, EMD$^\alpha$, and EMD, respectively, at anomaly detection on synthetic data. $|V| = 256$, scale-free exponent $\gamma = -2.3$. A series of 21 network states is generated using $\mathbb{P}_{nbr} = 0.1$ and $\mathbb{P}_{ext} = 0.01$ for normal and $\mathbb{P}_{nbr} = 0.08$ and $\mathbb{P}_{ext} = 0.03$ for anomalous network states' generation, respectively. The simulated anomalies are displayed as solid vertical lines.

We can see that, expectedly, SND(EMD)—the version of SND internally using the original EMD—does not recognize the extra mass (the newly activated users) in new network states, and, as a result, it considers all network states to be equivalent to the very first network state. SND(EMD$^\alpha$) reacts to the changes in the number of newly activated users in each network state, but it cannot distinguish between the cases when these new activations are expected (normal or in-network user activation) and when these new activations are anomalous (largely external user activation). As a result, SND(EMD$^\alpha$) does not recognize the anomalies, while SND = SND(EMD*) does by producing recognizable spikes at the times of anomalous network transitions.

6.5 Sensitivity to Opinion Dynamics Models

In the anomaly detection and user opinion prediction experiments, performance of SND stemmed from its being spatially-sensitive to the changes in the user opinion distribution, and promptly reacting to qualitative changes in the underlying opinion spread process. In this section, we conduct an experiment confirming that sensitivity of SND. We show the effectiveness of SND in detecting qualitative changes in the user opinions' evolution under an advanced opinion dynamics model, that cannot be spotted by the distance measures performing coordinate-wise comparison. To that end, we generate a number of pairs $\langle G(1), G(2)\rangle$ of network states adjacent in time ($G(2)$ is

generated from $G(1)$) over a synthetic scale-free network. Some of these pairs correspond to *normal transitions*, while others correspond to *anomalous transitions* in the network's evolution. For the normal transitions, $G(2)$ is generated from $G(1)$ using the Independent Cascade model [11]. For the anomalous transitions, most new user activations in $G(2)$ happen randomly, independently of the network's structure. We study the distances assigned to normal and anomalous network state transitions by SND and ℓ_1, and plot them as functions of the number n_Δ of users whose opinions change over each network state transition. The results are shown in Fig. 11.

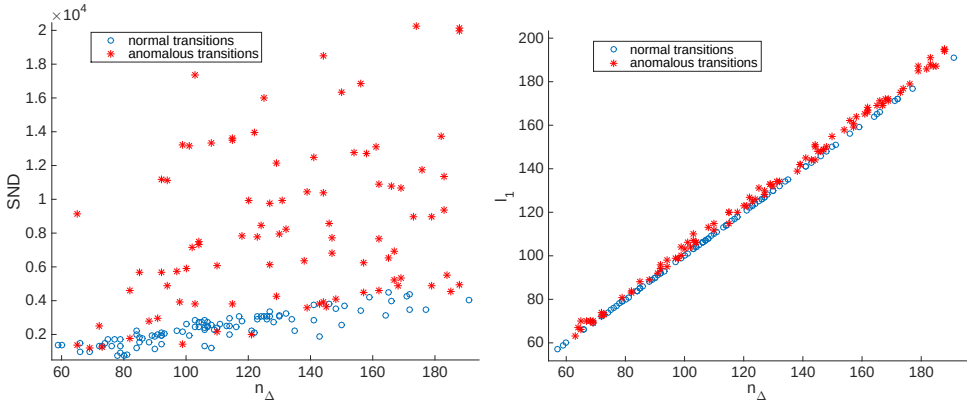


Fig. 11. SND and ℓ_1 distances between network states of normal (ICC) and anomalous (random) transitions.

We see that SND clearly separates anomalous transitions from normal ones, while ℓ_1 cannot discern anomalous network state transitions, as ℓ_1's value is mostly determined by n_Δ, which is representative of the distance measures performing coordinate-wise comparison.

6.6 Scalability of SND

We have implemented[10] SND in MATLAB and C++. We use the minimum-cost network flow solver CS2 [19] that implements Goldberg-Tarjan's algorithm [20], but, unlike it is prescribed by Theorem 5.3, does not use the two-edge push rule of Ahuja et al. [3]. Additionally, for computing shortest paths, our implementation of Dijkstra's algorithm uses a priority queue based on a binary heap, rather than a combination of a Fibonacci and a radix heaps [2]. As a result, our implementation of SND scales slightly worse than linearly—as guaranteed by Theorem 5.3—but still very well to be applicable to real-world social networks. Fig. 12 shows how our implementation of SND scales with respect to the number n of users in the network in comparison to a direct computation of SND using CPLEX' linear solver [16]. Our implementation's scalability with respect to the number n_Δ of users holding different opinions in two network states under comparison is shown in Fig. 13.

7 RELATED WORK

While the topic of understanding how opinions form and spread has been around for decades, with the rise of online social networks, the number of thematic works exploded—see surveys [1, 12, 43, 44] for reference. More recently, the analysis and modeling of polar opinion formation and spread has received attention [6]. The majority of the existing works, however, are dedicated to either designing opinion formation models, or designing algorithms for extremal problems over social networks, such as influence maximization.

[10] https://dynamo.cs.ucsb.edu/content/software/

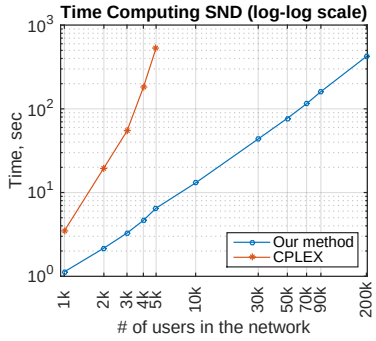


Fig. 12. Time for computing SND when the number of users having different opinion is $n_\Delta = 1000$, and the total number of users n in the network grows up 200k.

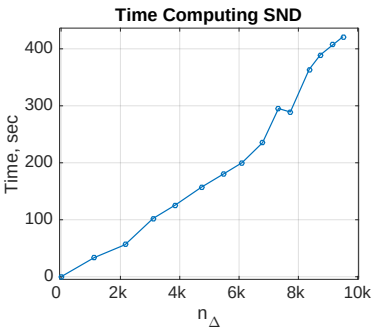


Fig. 13. Time for computing SND using our method, with the network size $n = 20$k, and the number n_Δ of users who changed their opinions growing up to 10k.

In this work, our algorithm design efforts target two specific applications—anomaly detection in the process of a network's state evolution and prediction of opinions of network users—that have not yet received enough attention. Below, in Sec. 7.1, we review existing works related to our two applications of choice. The methods we design in this work belong to the class of distance measure-based methods, and, hence, in Sec. 7.2, we also review a range of existing distance measures that can be considered competitors to SND when used with our anomaly detection and opinion prediction methods.

7.1 Anomaly Detection and Opinion Prediction

7.1.1 Anomaly Detection. There is a multitude of methods for anomaly detection in networks [5, 45], yet, most of them—such as [4] and [54]—are concerned with localizing an outlier part of the network. Detecting fake users in a social network is one representative application. We, however, are concerned with social network event detection—detecting moments of time when a social network behaves unexpectedly—the methods to address which are almost non-existent. Existing general methods for event detection, such as the ones based on tensor decomposition [45], would not scale to real-world large-scale social networks. The state of the art in large-scale network event detection is comparing network state snapshots using different *distance measures*. This is also our approach in this paper, and we provide a review of existing distance measures that can compete with ours in Sec. 7.2.

7.1.2 User Opinion Prediction. The majority of works concerned with social network user opinions target opinion detection from user-generated content, with only a few works' targeting the actual user opinion prediction. Conover et al. [15] detect communities in the network and assign user opinions based on community membership. Saito et al. [47] and De et al. [17] predict user opinions based on maximum-likelihood estimation, where predicted opinions are assigned based on those from the most likely state. Finally, Najar et al. [40] predict user opinions via simulating opinion formation models and tracking which opinions are assigned to the target users in simulation. We use the mentioned approaches as our opinion prediction baselines in Sec. 6.3.

7.2 Distance Measures

There is a large number of existing distance measures used in vector spaces, including ℓ_p, Hamming, Canberra, Cosine, Kullback-Leibler, and Quadratic Form [23] distances. However, none of them are

suitable for the comparison of social network states, since these distance measures either compare opinion vectors coordinate-wise, thereby, not capturing user interaction in the network, or in the case of Quadratic Form distance, capture the user interaction in a very limited way, being unaware of the underlying process that causes the difference between two user opinion distributions.

Existing graph-oriented distance measures are also unsuitable for comparing network states with polar opinions. The first type of such distance measures is graph isomorphism-based distance measures, such as *largest common subgraph* [10]. They are node state-oblivious, and, hence are not applicable to the comparison of network states. Another type of graph distance measures is *Graph Edit Distance (GED)-based* distance measures [18] that define the distance between two networks as the cost of an optimal sequence of node or edge insertions, deletions, and substitutions, transforming one network into another. GED can be node state-aware, but its value is not interpretable from the opinion dynamics point of view, and even its approximate computation takes cubic time (a single computation of GED on a 10k-node network on our hardware takes about a month). DeltaCon [28] is a scalable graph-oriented distance measure, yet, it quantifies the networks' structural difference, while we focus on node states.

A third class of distance measures includes *iterative distance measures* [8, 26, 29, 38], which express similarity of the nodes of two networks recursively, use a fix-point iteration to compute node similarities, and, then, aggregate node similarities to obtain the similarity of two networks. These share the problem of GED—they do not capture the way competing opinions spread in the network. The same drawback is shared by the related diffusion-based distance measures [24, 32, 48], that compare network states by quantifying how differently a heat diffusion process proceeds in the network when the network states defined the initial temperatures of the nodes.

The last class includes *feature-based* distance measures [7, 52, 53, 53], which compare either the distributions of local node properties (e.g., degree, clustering coefficient) or the spectra of two networks. Despite their efficient computability, such distance measures do not fit the comparison of network states with polar opinions. The spectral distance measures are inadequate because they do not deal with node states directly[11], while other feature-based distance measures only deal with summaries based on opinion of each kind, thus, being unable to capture competition of opinions.

8 LIMITATIONS

Despite the demonstrated effectiveness and efficiency of SND, there are scenarios in which its use is either prohibitively or unnecessarily expensive.

- *When SND is too expensive:* One reason to choose a simpler distance measure, such as ℓ_p, over SND is the latter's computational cost. While it is asymptotically linear in the number n of nodes, its cost can potentially be too high in practice for networks having 100M+ nodes. In such networks, a single computation of SND can take several days. If it is nonetheless desirable to use SND on such a large network, one can partition that network into clusters of tractable size and perform the SND-based analysis on each individual cluster.
- *When SND is unnecessarily expensive:* Using SND may be superfluous if the changes in the rate of new user activation reveal enough information for the target application. For example, the user activation rate alone is clearly enough to detect a presidential election day. For detection of such "anomalies", a distance measure as simple as Hamming distance may suffice.

9 FUTURE RESEARCH

Among the directions for future research are the following.

[11] Even if node states are artificially encoded into a network's structure, there is still a possibility for two structurally different networks to have identical spectra and, hence, a zero spectral distance.

- *New applications:* Since SND is, effectively, the first distance measure designed specifically for the comparison of states of a social network containing competing opinions, one potential future research direction is using SND in other applications operating in a metric space setting, such as network state classification, clustering, and search.
- *Combining macro-level distance measure-based and user-level analysis:* It may be lucrative to combine SND with non-distance measure-based methods. For example, in the method of Conover et al. [15] that predicts opinions based on the content of the users' tweets, the objective function can be augmented with an SND-based term, thereby, performing opinion fitting at both the micro-level of each user and the macro-level of the entire network.
- *Design of a distance measure for both structural and opinion changes:* Finally, it may be fruitful to design a distance measure that would capture changes in both the opinions of the users and the structure of the social network simultaneously. Such a distance measure would be more computationally complex than SND, but would result in a more accurate analysis when the network structure changes a lot from network state to network state.

10 CONCLUSION

In this paper, we proposed Social Network Distance (SND)—the first distance measure for comparing the states of a social network containing polar opinions. Our distance measure quantifies how likely one state of a social network has evolved into another state under a given model of opinion dynamics. Despite the high computational complexity of the transportation problem underlying SND, we propose a linear-time algorithm for its exact computation, making SND applicable to real-world online social networks. We demonstrate the usefulness of SND in detecting anomalous network states and predicting user opinions, where SND-based methods consistently outperform competitors. Our anomaly detection method achieves a true positive rate (TPR) of 0.83 when the false positive rate (FPR) is 0.3, while the next best method's TPR is only 0.4 at the same FPR. The accuracy of SND-based method for user opinion prediction in Twitter data is 75.63%, which is 7.5% higher than that of the next best method. We also show that, unlike the distance measures performing coordinate-wise comparison, SND can detect qualitative changes in the network's evolution pattern. SND is a powerful alternative to simpler distance measures, and is effective when such summaries of network users' behavior as the number of active users are uninformative, and a deeper insight into the opinion dynamics process is required.

A APPENDIX

A.1 Opinion Acquisition Under Independent Cascade and Linear Threshold Models

In this section, we provide two examples of how to define the likelihood $\mathbb{P}_{ij}(P, v)$ of user j adopting opinion v from user i in network state P for the versions of Independent Cascade and Linear Threshold models supporting competing opinions.

Independent Cascade: For the version of the Independent Cascade model [11] supporting multiple opinion values, $\mathbb{P}_{ij}(P, v)$ is defined as follows

$$\mathbb{P}^{\mathrm{IC}}_{ij}(P, v) = \begin{cases} 0 & \text{if } d_j(\{i\}) > d_j(I_P^{+1} \cup I_P^{-1}), \\ 1 & \text{else if } P_i = v \text{ and } P_j = v, \\ \frac{\max(0, p_{ij} - \varepsilon)}{\sum\limits_{i \in act(P,j)} p_{ij}} & \text{else if } P_i = v \text{ and } P_j = 0, \\ \varepsilon & \text{otherwise,} \end{cases}$$

where I_P^v is a set of users holding opinion v in network state P, p_{uv} is an edge activation probability [22], $d_j(I)$ is the length of the shortest path from users I to user j, $act(P, j) = \{k \mid k \in$

$I_P^{+1} \cup I_P^{-1}$ and $d_j(\{k\}) = d_j(I_P^{+1} \cup I_P^{-1})\}$ is the set of users active in network state P closest to user j, and ε is a negligible likelihood of an "impossible" event.

In the original model, $\varepsilon = 0$, that is, neutral users cannot infect others, and active users neither drop their opinions nor spread opinions opposite to their own. That, however, would lead to the distances between many network states to be $+\infty$ since the opinion evolution may follow the assumed opinion dynamics model *not exactly*. In order to obviate this issue, we aim to—instead of just declaring two network states qualitatively unreachable—to always quantify the distance between them, and, thus, assign some negligible probabilities ε to the events that original opinion dynamics models posit as impossible.

Linear Threshold: For the version of Linear Threshold model [9] supporting multiple opinion values, $\mathbb{P}_{ij}(P, v)$ is defined as

$$\mathbb{P}^{\mathrm{LT}}_{ij}(P, v) = \begin{cases} 0 & \text{if } i \notin N^{in}(P, j), \\ 1 & \text{else if } P_i = v \wedge P_j = j, \\ \frac{(1-\varepsilon)\omega_{ij}}{\sum_{k \in N^{in}(P,j)} \omega_{kj}} & \text{else if } P_i = v \text{ and } P_j = 0 \text{ and } \sum_{k \in N^{in}(P,j)} \omega_{kj} \geq \theta_j, \\ \varepsilon & \text{otherwise,} \end{cases}$$

where ω_{ij} is an edge weight reflecting relative influence of i upon j, θ_i is user i's opinion switching threshold, $N^{in}(P, j)$ is the set of j's in-neighbors active in network state P, and ε has the same semantics of a negligible likelihood of an "impossible" event as in the earlier case of the Independent Cascade model.

A.2 Equivalence of EMD$^{\alpha}$ and $\widehat{\mathrm{EMD}}$

In Sec. 4.1, we used EMD$^{\alpha}$—an existing version of Earth Mover's Distance—and relied upon its equivalence to another existing version of EMD, namely,

$$\widehat{\mathrm{EMD}}(P, Q, D) = \mathrm{EMD}(P, Q, D) \min\left\{\sum_{i=1}^{n} P_i, \sum_{j=1}^{n} Q_j\right\} + \alpha \max_{i,j}\{D_{ij}\} \left|\sum_{i=1}^{n} P_i - \sum_{j=1}^{n} Q_j\right|.$$

In this section, we provide a formal definition of EMD$^{\alpha}$, and establish its equivalence with $\widehat{\mathrm{EMD}}$ in Theorem A.1. Formally, EMD$^{\alpha}$ is defined as follows.

$$P = [P_1, \ldots, P_n],\ Q = [Q_1, \ldots, Q_n],$$

$$P_{bank} = \sum_{j=1}^{n} Q_j, \quad \widetilde{P} = [P_1, \ldots, P_n, P_{bank}],$$

$$Q_{bank} = \sum_{i=1}^{n} P_i, \quad \widetilde{Q} = [Q_1, \ldots, Q_n, Q_{bank}],$$

$$\widetilde{D} = \left[\begin{array}{c|c} D_{n\times n} & \begin{array}{c} | \\ \alpha \max_{i,j}\{D_{ij}\} \\ | \end{array} \\ \hline -\ \alpha \max_{i,j}\{D_{ij}\}\ - & 0 \end{array}\right],$$

$$\mathrm{EMD}^{\alpha}(P, Q) = \mathrm{EMD}(\widetilde{P}, \widetilde{Q}, \widetilde{D}) \cdot \left(\sum_{i=1}^{n} P_i + \sum_{j=1}^{n} Q_j\right).$$

Despite the syntactic differences between EMD^{α} and $\widehat{\text{EMD}}$, the following Theorem establishes that they are actually numerically equivalent.

Theorem A.1. *If ground distance $D \in \mathbb{R}^{n\times n}$ is metric, and $\alpha \geq \frac{1}{2}$, so that both EMD^{α} and $\widehat{EMD}$ are metric [34, 41], then $\forall P, Q \in \mathbb{R}^{+n}$: $EMD^{\alpha}(P, Q, D) = \widehat{EMD}(P, Q, D)$.*

Proof. Without loss of generality, let us assume that $\sum P_i \leq \sum Q_j$, and use the following notation

$$\Delta = \Delta(P, Q) = \Big| \sum_{i=1}^{n} P_i - \sum_{j=1}^{n} Q_j \Big|, \quad \gamma = \alpha \max_{i,j} \{D_{ij}\},$$

so that the expression for $\widehat{\text{EMD}}$ gets rewritten as

$$\widehat{\text{EMD}}(P, Q) = \text{EMD}(P, Q) \min \Big\{ \sum_{i=1}^{n} P_i, \sum_{j=1}^{n} Q_j \Big\} + \gamma\Delta.$$

Our aim is to show that EMD^{α} has exactly the same expression as $\widehat{\text{EMD}}$ as long as they both are metric. To do so, let is consider how a unit of mass can be transported from network state P to network state Q, both extended with a bank node, as per the definition of EMD^{α}.

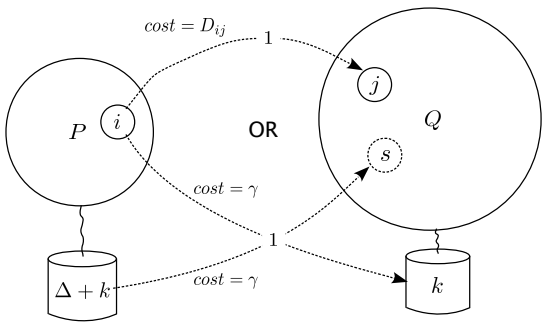


Fig. 14. Two qualitatively different ways to transport a unit of mass from extended network state $\widetilde{P} = [P_1, \ldots, P_n, k + \Delta]$ to extended network state $\widetilde{Q} = [Q_1, \ldots, Q_n, k]$, where $\sum P_i \leq \sum Q_j$. Dashed arrows represent the flow of mass. The bank node is attached to every node of each network state. $k = \sum P_i$, so that the total masses of two extended network states are equal.

As shown in Fig. 14, there are two qualitatively different alternatives for moving a unit of mass from regular (non-bank) node i of network state $\widetilde{P}$: a unit of mass can be moved either to a regular node j or to the bank node of $\widetilde{Q}$.

In the first case, the total transportation cost of a unit of mass is exactly the ground distance $\widetilde{D}_{ij} = D_{ij}$ between regular nodes i and j.

In the second case, the immediate cost of transportation to the bank node is $\widetilde{D}_{i,bank} = \gamma$. However, because we have routed mass from a regular node to the bank node, there exists a regular node s in $\widetilde{Q}$ having a "mass deficit" that has to be fulfilled from the bank node of $\widetilde{P}$. Thus, if we move a unit of mass from a regular node of $\widetilde{P}$ to the bank node of $\widetilde{Q}$, there is an additional incurred cost γ of moving an additional unit of mass from the bank node of $\widetilde{P}$ to some regular node of $\widetilde{Q}$. Hence, the total cost of transportation of a unit of mass in the second case is

$$\gamma + \gamma = 2\alpha \max_{i,j} D_{ij} \geq (\text{since } \alpha \geq 0.5) \geq \max_{i,j} D_{ij}.$$

Thus, from the point of view of optimal mass transportation, it may never be preferable to move a unit of mass from a regular node to the bank node if there is an option to transport mass from a regular node to another regular node. Consequently, an optimal solution to the EMD$^\alpha$'s transportation problem can be decomposed as follows.

$$\begin{aligned}
\mathrm{EMD}^\alpha(P,Q,D) &= \mathrm{EMD}(\widetilde{P},\widetilde{Q},\widetilde{D})\left(\sum_{i=1}^{n} P_i + \sum_{j=1}^{n} Q_j\right) \\
&= \min_{\{f_{ij}\}} \sum_{i,j=1}^{n+1} f_{ij}\widetilde{D}_{ij} = (\text{let } b = n+1) \\
&= \min_{\{f_{ij}\}} \Bigg[\underbrace{\sum_{i,j=1}^{n} f_{ij}\widetilde{D}_{ij}}_{\substack{\text{regular nodes}\\ \text{to regular nodes}}} + \underbrace{\sum_{i=1}^{n} f_{ib}\widetilde{D}_{ib}}_{\substack{\text{regular nodes}\\ \text{to bank node}}} + \underbrace{\sum_{j=1}^{n} f_{bj}\widetilde{D}_{bj}}_{\substack{\text{bank node to}\\ \text{regular nodes}}} + \underbrace{f_{bb}\widetilde{D}_{bb}}_{\substack{\text{bank node}\\ \text{to bank node}}} \Bigg] \\
&= \min_{\{f_{ij}\}} \Bigg[\sum_{i,j=1}^{n} f_{ij}D_{ij} + \gamma \underbrace{\sum_{j=1}^{n} f_{bj}}_{\Delta} \Bigg] = \min_{\{f_{ij}\}} \Bigg[\sum_{i,j=1}^{n} f_{ij}D_{ij} + \gamma\Delta \Bigg] \\
&= \min_{\{f_{ij}\}} \Bigg[\sum_{i,j=1}^{n} f_{ij}D_{ij} \Bigg] + \gamma\Delta = \mathrm{EMD}(P,Q,D) \times \min\left\{\sum P_i, \sum Q_j\right\} + \gamma\Delta \\
&= \widehat{\mathrm{EMD}}(P,Q,D).
\end{aligned}$$

An additional useful observation, formalized below as Corollary A.2—that will subsequently show instrumental in the proof of Theorem 4.2—is that a particular value of k does not matter for EMD$^\alpha$, since in every optimal solution of its underlying transportation problem, any amount of mass exceeding Δ in the bank node of the lighter network state is transported at cost $\widetilde{D}_{bank,bank} = 0$ to the bank node of the heavier network state. □

Corollary A.2. *For network states $P = [P_1, \ldots, P_n]$ and $Q = [Q_1, \ldots, Q_n]$, and ground distance D, if $\sum P_i = \sum Q_j$ and D is metric, then for all $k \geq 0 \in \mathbb{R}^+$, the following holds.*

$$EMD\left([P,k],[Q,k],\left[\begin{array}{c|c} D & \begin{array}{c}|\\ \omega \\ |\end{array} \\ \hline -\,\omega\,- & 0 \end{array}\right]\right) = EMD(P,Q,D),$$

where $[X,k]$ is network state X extended with a single bank node with value k and a uniformly defined ground distance $\omega \geq \frac{1}{2}\max_{i,j} D_{i,j}$ to/from the regular nodes of X.

Informally, Corollary A.2 states that, for any two network states of equal total mass, we can increase their total masses by an arbitrary value without affecting the EMD between them.

A.3 Proof of Theorem 4.2

Proof (Theorem 4.2). Let us define $M = \max_{X \in \mathcal{H}} \sum_k X_k < \infty$. Next, we define an auxiliary distance measure EMD' as follows.

$$\mathrm{EMD}'(P, Q, D) = \mathrm{EMD}(P', Q', D'),$$

$$P' = [\widetilde{P}, M - \sum_i \widetilde{P}_i],$$

$$Q' = [\widetilde{Q}, M - \sum_j \widetilde{Q}_j],$$

$$D' = \left[\begin{array}{c|c} \widetilde{D} & \begin{array}{c} | \\ \max\limits_{i,j} \{\widetilde{D}_{ij}\}/2 \\ | \end{array} \\ \hline - \max\limits_{i,j} \{\widetilde{D}_{ij}\}/2 - & 0 \end{array}\right],$$

where $\widetilde{P}$, $\widetilde{Q}$, and $\widetilde{D}$ are the extended network states, and the extended ground distance, respectively, as defined by EMD*. From the definition of EMD* (10), it follows that $\sum \widetilde{P}_i = \sum \widetilde{Q}_i$ and, hence $M - \sum \widetilde{P}_i = M - \sum \widetilde{Q}_i = k$. Thus, since $\sum P'_i = \sum Q'_i = M$, D is metric, and $k \geq 0$, from Corollary A.2 with $P = \widetilde{P}$, $Q = \widetilde{Q}$, $D = \widetilde{D}$, and $\omega = \frac{1}{2} \max\limits_{i,j} \widetilde{D}_{ij}$, we have

$$\mathrm{EMD}'(P, Q, D) = \mathrm{EMD}(P', Q', D') = (\text{from Corollary A.2}) = \mathrm{EMD}(\widetilde{P}, \widetilde{Q}, \widetilde{D}) =$$

$$= (\text{from definition (10) of EMD}^\star) = \frac{\mathrm{EMD}^\star(P, Q, D)}{\max\left\{\sum P_i, \sum Q_j\right\}}.$$

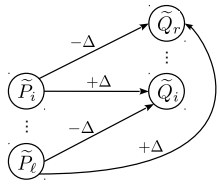
Thus, EMD* is metric iff EMD$'$ is metric. The latter's metricity, according to Theorem 2.1, requires equality of total masses of all network states and metricity of the ground distance. From the definition of EMD$'$, it is clear that *all* network states P' and Q' supplied to EMD by EMD$'$ have the same total mass M. As to metricity of the ground distance, the identity of indiscernibles and symmetry straightforwardly follow from the corresponding properties of the original ground distance D and our choice of the ground distances to/from the bank nodes to be non-negative and symmetric. The triangle inequality trivially holds as network state extension does not introduce any new triangles. Hence, by Theorem 2.1, EMD$'$ and, consequently, EMD* is metric. □

A.4 Proof of Reduction Lemma 5.2

Proof (Lemma 5.2). First, we will show that there is always an optimal plan f_{ij} in the problem of optimal mass transportation from $\widetilde{P}$ to $\widetilde{Q}$ over $\widetilde{D}$ such that $\forall 1 \leq i \leq n : f_{ii} = \min\{\widetilde{P}_i, \widetilde{Q}_i\} = M$, and, then, use such a plan to argue about the value of EMD*.

Consider an arbitrary optimal transportation plan $\hat{f}_{ij}$, and assume that $\exists i \in [1; n] : \delta = M - \hat{f}_{ii} > 0$. We will now use $\hat{f}$ to construct another optimal transportation plan $f^\dagger_{ij}$ such that $f^\dagger_{ii} = M$. Initially, we put $f^\dagger = \hat{f}$ and, then, re-route mass flows in $f^\dagger$ to eventually achieve the desired value of $f^\dagger_{ii}$.

Since, initially, $f^\dagger_{ii} < M$, the remaining at least δ units of mass should be distributed by $\widetilde{P}_i$ and consumed by $\widetilde{Q}_i$ to/from other consumers/suppliers. Among those, let us pick the ones that supply/consume the least amount of mass to $\widetilde{Q}_i$ and from $\widetilde{P}_i$, respectively: $\ell = \arg\min_{j \neq i} f^\dagger_{ji}$, and $r = \arg\min_{j \neq i} f^\dagger_{ij}$. Without loss of generality, let us assume that $f^\dagger_{\ell i} \leq f^\dagger_{ir}$ and denote $\Delta = \min\{f^\dagger_{\ell i}, \delta\}$. Now, we will re-route Δ units of mass in $f^\dagger$ as follows.



$$f^\dagger_{\ell i} \leftarrow f^\dagger_{\ell i} - \Delta, \quad f^\dagger_{\ell r} \leftarrow f^\dagger_{\ell r} + \Delta, \quad f^\dagger_{ir} \leftarrow f^\dagger_{ir} - \Delta, \quad f^\dagger_{ii} \leftarrow f^\dagger_{ii} + \Delta.$$

The updated transportation plan is legal, as the total amount of mass supplied or consumed by each node has not changed. The total cost of $f^\dagger$ has been updated as follows

$$\begin{aligned} newcost(f^\dagger) \leftarrow\ & cost(f^\dagger) - \Delta\widetilde{D}_{\ell i} - \Delta\widetilde{D}_{ir} + \Delta\widetilde{D}_{ii} + \Delta\widetilde{D}_{\ell r} = \\ &= \big(\text{since } D \text{ and, hence, } \widetilde{D} \text{ is semimetric, } \widetilde{D}_{ii} = 0\big) = cost(f^\dagger) - \Delta(\widetilde{D}_{\ell i} + \widetilde{D}_{ir} - \widetilde{D}_{\ell r}) \\ &\le \big(\text{since } \widetilde{D} \text{ is semimetric, } \widetilde{D}_{\ell i} + \widetilde{D}_{ir} \ge \widetilde{D}_{\ell r}\big) \le cost(f^\dagger). \end{aligned}$$

Since the cost of the obtained legal plan $f^\dagger$ cannot be strictly less than the cost of an optimal plan $\hat{f}$, the performed update of $f^\dagger$ has not changed its cost, and the updated $f^\dagger$ is still an optimal plan. The described above re-routing procedure is repeatedly performed on $f^\dagger$ until $f^\dagger_{ii}$ reaches $M = \min\{\widetilde{P}_i, \widetilde{Q}_i\}$.

Finally, to see why the statement of the lemma holds, we observe that the value of EMD* is the cost of any optimal transportation plan, and the cost of $f^\dagger$ in particular. However, the cost of $f^\dagger$ does not depend on $f^\dagger_{ii}$, since, due to semimetricity of $\widetilde{D}$, mass $f^\dagger_{ii}$ gets transported at cost $\widetilde{D}_{ii} = 0$. Thus, M can be subtracted from $\widetilde{P}_i$, $\widetilde{Q}_i$, and $f^\dagger_{ii}$, without affecting the total cost of $f^\dagger$. The solution of the latter modified transportation problem, however, is exactly

$$\text{EMD}^\star([P_1, \ldots, P_{i-1}, P_i - M, P_{i+1}, \ldots, P_n], [Q_1, \ldots, Q_{i-1}, Q_i - M, Q_{i+1}, \ldots, Q_n], D).$$

□

A.5 Example of Computing SND

In this section, we provide an example of computing SND over a toy network, and discuss the semantics of SND's value. Consider the following network G on 6 nodes, with $\{0, 1\}$-adjacency matrix A_{01}, and three states X, Y, and Z of that network.

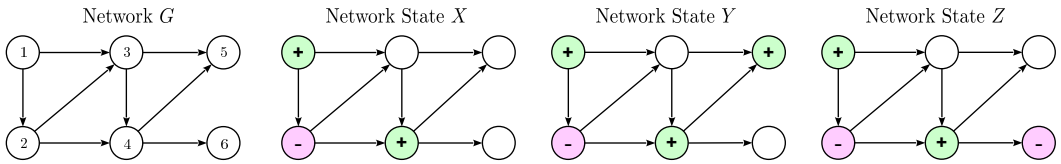


Fig. 15. Example network G and three network states X, Y, and Z. Pluses correspond to node values +1, and minuses—to −1.

$$A_{01} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{aligned} X &= [+1, -1, \quad 0, +1, \quad 0, \quad 0]^\intercal \\ Y &= [+1, -1, \quad 0, +1, +1, \quad 0]^\intercal \\ Z &= [+1, -1, \quad 0, +1, \quad 0, -1]^\intercal \end{aligned}$$

In this example, SND will rely on the following underlying opinion propagation model, defined via $\mathbb{P}_{ij}(X,+1)$ and $\mathbb{P}_{ij}(X,-1)$—the likelihoods of positive and negative opinion spread between pairs of nodes in network state X:

$$\mathbb{P}(X,+1) = \begin{bmatrix} \varepsilon & 0.01 & 0.8 & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & 0.2 & 0.2 & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & 0.01 & 0.5 \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \end{bmatrix}, \quad \mathbb{P}(X,-1) = \begin{bmatrix} \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & 0.5 & 0.01 & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & 0.01 & 0.2 & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & 0.01 & 0.01 \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \end{bmatrix},$$

where $\varepsilon = 0.001$ is the likelihood of an "implausible event"—discussed in Appendix A.1—which makes sure we can always quantify distance between network states, even if some opinion adoptions seem implausible, possibly, due to missing data. We will not observe such implausible events when comparing X and Y, but the comparison of X and Z will be affected by the choice of ε.

In what follows, we will show how to compute an asymmetric version of SND

$$\mathrm{SND}^{asym}(X,Y) = \sum_{v\in\{-1,+1\}} \mathrm{EMD}^{\star}(X^v, Y^v, D(X,v))$$

and compare the obtained value with that of $\mathrm{SND}^{asym}(X,Z)$. Computation of full $\mathrm{SND}(X,Y) = \frac{1}{2}[\mathrm{SND}^{asym}(X,Y) + \mathrm{SND}^{asym}(Y,X)]$ would be performed using the same procedure.

Expression for $\mathrm{SND}^{asym}(X,Y)$ contains a sum with two terms. For the term corresponding to $v=-1$, we have

$$X^{-1} = [0,1,0,0,0,0]^\intercal = Y^{-1},$$

so $\mathrm{EMD}^{\star}(X^{-1}, Y^{-1}, D(X,-1)) = 0$. Thus, we need to compute only $\mathrm{EMD}^{\star}(X^{+1}, Y^{+1}, D(X,+1))$:

$$X^{+1} = [1,0,0,1,0,0]^\intercal \quad Y^{+1} = [1,0,0,1,1,0]^\intercal$$

The obtained X^{+1} and Y^{+1} have different masses

$$\Delta = \Big|\sum_{i=1}^{n} X_i^{+1} - \sum_{i=1}^{n} Y_i^{+1}\Big| = |2-3| = 1,$$

so $\mathrm{EMD}^{\star}$ extends X^{+1} and Y^{+1}, equalizing their masses, as follows.

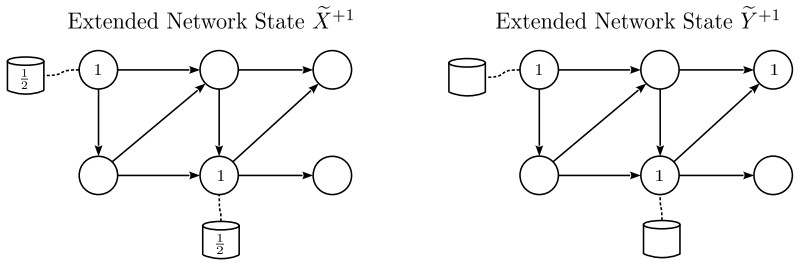


Fig. 16. X^{+1} and Y^{+1} extended with bank nodes. Blank nodes (regular or bank) have zero values. The bank nodes attached to nodes 1 and 4 are indexed as 7 and 8, respectively.

$$\widetilde{X}^{+1} = [(X^{+1})^\intercal \mid 0.5, 0.5]^\intercal = [1,0,0,1,0,0,0.5,0.5]^\intercal,$$
$$\widetilde{Y}^{+1} = [(Y^{+1})^\intercal \mid 0.0, 0.0]^\intercal = [1,0,0,1,1,0,0,0]^\intercal.$$

The corresponding extended ground distance matrix $D(X,+1)$ is obtained as follows:

$$\widetilde{D}(X,+1) = a2asp(-\log(\widetilde{\mathbb{P}}(X,+1))),$$

$$\widetilde{\mathbb{P}}(X,+1) = \left[\begin{array}{c|cc} \mathbb{P}(X,+1) & \begin{array}{c}|\\ \exp\{-\gamma\}I_1\\ |\end{array} & \begin{array}{c}|\\ \exp\{-\gamma\}I_4\\ |\end{array} \\ \hline \begin{array}{c}-\ \exp\{-\gamma\}I_1\ -\\ -\ \exp\{-\gamma\}I_4\ -\end{array} & \multicolumn{2}{c}{\mathbb{0}_{2\times 2}} \end{array}\right]$$

where I_j is the j'th column of the identity matrix, $\gamma = 0.001$ is the log-likelihood assigned to the edges to/from the added bank nodes, and $a2asp(A)$ is the matrix of the lengths of all-to-all shortest paths in the network with adjacency matrix A. In an efficient computation of EMD*, outlined in Theorem 5.3, we would *not* need to compute all shortest paths, but in this example we do so for the sake of simplicity.

Finally, we compute $\text{EMD}^\star(X^{+1}, Y^{+1}, D(X,+1))$ as the value of a min-cost bipartite network flow with sources $\widetilde{X}^{+1}$, destinations $\widetilde{Y}^{+1}$, and transportation costs $\widetilde{D}(X,+1)$:

$$\text{EMD}^\star(X^{+1}, Y^{+1}, D(X,+1)) = mcf(\widetilde{X}^{+1}, \widetilde{Y}^{+1}, \widetilde{D}(X,+1)) \approx 4.2.$$

The corresponding optimal flow is shown in Fig. 17. Thus, we conclude that $\text{SND}^{asym}(X,Y) \approx 4.2$.

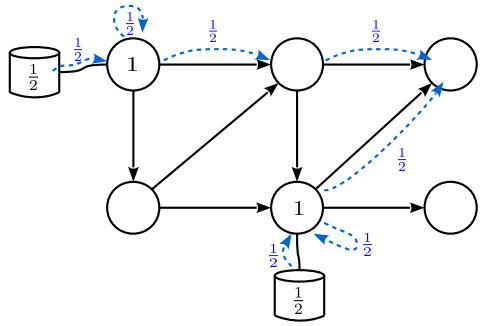


Fig. 17. Optimal flow of opinion +1 from sources $\widetilde{X}^{+1}$ to destinations $\widetilde{Y}^{+1}$ through network in state $\widetilde{X}^{+1}$. The cost of this flow is ≈ 4.2, where node-to-node transportation costs are defined via $\widetilde{D}(X,+1)$.

If we translate the obtained distance value $\text{SND}^{asym}(X,Y) = 4.2$ from log-scale, we will get $\exp\{-\text{SND}^{asym}(X,Y)\} = 0.015$. Taking into account that SND quantifies the likelihood of a network's state transition, does the value of 0.015 make sense on its own as a probability? If we assume that, in reality, node 5 could have got opinion +1 through either of two equiprobable paths $1 \longrightarrow 3 \longrightarrow 5$ or $4 \longrightarrow 5$, then the likelihood of the network's transitioning from X^{+1} to Y^{+1} would have been

$$0.5 \cdot \mathbb{P}_{13}(X,+1) \cdot \mathbb{P}_{35}(X,+1) + 0.5 \cdot \mathbb{P}_{45}(X,+1)$$
$$= 0.5 \cdot 0.8 \cdot 0.5 + 0.5 \cdot 0.001 \approx 0.20 \gg 0.015 = \exp\{-\text{SND}^{asym}(X,Y)\}.$$

We see that SND may considerably underestimate the likelihood of network state transition, and, hence, it may not be reasonable to interpret SND's value *in isolation* as a (log-)likelihood. However, the value of SND between two network states makes sense when used in relation to other values

of SND between other network state pairs. For example, if we repeated the above procedure to compute $\mathrm{SND}^{asym}(X, Z)$, we would have obtained

$$\mathrm{SND}^{asym}(X, Z) \approx 7.9 \gg 4.2 = \mathrm{SND}^{asym}(X, Y),$$

which is expected, as transition from X to Z includes propagating negative opinion through node 4 holding positive opinion, while transition from X to Y does not involve such unlikely opinion adoptions. SND effectively captures such a difference, indicating that Y is a more likely successor to X than Z.

ACKNOWLEDGMENTS

This work was supported by the U. S. Army Research Laboratory and the U. S. Army Research Office under grant number W911NF-15-1-0577, and the National Science Foundation under grant number IIS-1817046.

REFERENCES

[1] Daron Acemoglu and Asuman Ozdaglar. 2011. Opinion dynamics and learning in social networks. *Dynamic Games and Applications* 1, 1 (2011), 3–49.
[2] Ravindra K Ahuja, Kurt Mehlhorn, James Orlin, and Robert E Tarjan. 1990. Faster algorithms for the shortest path problem. *J. ACM* 37, 2 (1990), 213–223.
[3] Ravindra K Ahuja, James B Orlin, Clifford Stein, and Robert E Tarjan. 1994. Improved algorithms for bipartite network flow. *SIAM J. Comput.* 23, 5 (1994), 906–933.
[4] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. 2010. Oddball: Spotting anomalies in weighted graphs. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 410–421.
[5] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery* 29, 3 (2015), 626–688.
[6] Victor Amelkin, Francesco Bullo, and Ambuj K Singh. 2017. Polar opinion dynamics in social networks. *IEEE Trans. Automat. Control* 62, 11 (2017), 5650–5665.
[7] Michele Berlingerio, Danai Koutra, Tina Eliassi-Rad, and Christos Faloutsos. 2013. Network similarity via multiple social theories. In *Proc. of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 1439–1440.
[8] Vincent D Blondel, Anahí Gajardo, Maureen Heymans, Pierre Senellart, and Paul Van Dooren. 2004. A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM Rev.* 46, 4 (2004), 647–666.
[9] Allan Borodin, Yuval Filmus, and Joel Oren. 2010. Threshold models for competitive influence in social networks. In *Proc. of International Workshop on Internet and Network Economics (WINE)*. Springer, 539–550.
[10] Horst Bunke and Kim Shearer. 1998. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters* 19, 3 (1998), 255–259.
[11] Tim Carnes, Chandrashekhar Nagarajan, Stefan M Wild, and Anke Van Zuylen. 2007. Maximizing influence in a competitive social network: A follower's perspective. In *Proc. of ACM International Conference on Electronic Commerce (EC)*. 351–360.
[12] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. 2009. Statistical physics of social dynamics. *Reviews of Modern Physics* 81, 2 (2009), 591.
[13] Kenneth L Clarkson. 2006. Nearest-neighbor searching and metric space dimensions. In *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*, Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk (Eds.). The MIT Press.
[14] Raviv Cohen and Derek Ruths. 2013. Classifying Political Orientation on Twitter: It's Not Easy!. In *Proc. of AAAI Conference on Web and Social Media (ICWSM)*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6128/6347
[15] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of Twitter users. In *Proc. of IEEE International Conference on Social Computing and Networking (SocialCom)*. 192–199.
[16] IBM ILOG CPLEX. 2009. V12. 1: User's Manual for CPLEX. *International Business Machines Corp.* 46, 53 (2009), 157.
[17] Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, and Manuel Gomez Rodriguez. 2016. Learning and forecasting opinion dynamics in social networks. In *Advances in Neural Information Processing Systems*. 397–405.

[18] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. 2010. A survey of Graph Edit Distance. *Pattern Analysis and Applications* 13, 1 (2010), 113–129.
[19] Andrew Goldberg. 1997. An efficient implementation of a scaling minimum-cost flow algorithm. *Journal of Algorithms* 22, 1 (1997), 1–29.
[20] Andrew Goldberg and Robert Tarjan. 1987. Solving minimum-cost flow problems by successive approximation. In *Proc. of ACM Symposium on Theory of Computing (STOC)*. 7–18.
[21] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. 2010. Inferring networks of diffusion and influence. In *Proc. of ACM Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*. 1019–1028.
[22] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proc. of ACM International Conference on Web Search and Data Mining (WSDM)*. 241–250.
[23] James Hafner, Harpreet S. Sawhney, William Equitz, Myron Flickner, and Wayne Niblack. 1995. Efficient color histogram indexing for quadratic form distance functions. *Pattern Analysis and Machine Intelligence* 17, 7 (1995), 729–736.
[24] David K Hammond, Yaniv Gur, and Chris R Johnson. 2013. Graph diffusion distance: A difference measure for weighted graphs based on the graph Laplacian exponential kernel. In *Proc. of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 419–422.
[25] Frederick S Hillier and Gerald J Lieberman. 1995. Introduction to Mathematical Programming. (1995).
[26] Glen Jeh and Jennifer Widom. 2002. SimRank: A measure of structural-context similarity. In *Proc. of ACM Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*. 538–543.
[27] Narendra Karmarkar. 1984. A new polynomial-time algorithm for linear programming. In *Proc. of ACM Symposium on Theory of Computing (STOC)*. 302–311.
[28] Danai Koutra, Joshua T Vogelstein, and Christos Faloutsos. 2013. DeltaCon: A principled massive-graph similarity function. In *Proc. of SIAM International Conference on Data Mining (SDM)*. SIAM, 162–170.
[29] Elizabeth A Leicht, Petter Holme, and Mark E J Newman. 2006. Vertex similarity in networks. *Physical Review E* 73, 2 (2006), 026120.
[30] Sanjiva K Lele. 1992. Compact finite difference schemes with spectral-like resolution. *J. Comput. Phys.* 103, 1 (1992), 16–42.
[31] Longjie Li, Min Ma, Peng Lei, Xiaoping Wang, and Xiaoyun Chen. 2014. A linear approximate algorithm for Earth Mover's Distance with thresholded ground distance. *Mathematical Problems in Engineering* (2014).
[32] Haibin Ling and Kazunori Okada. 2006. Diffusion distance for histogram comparison. In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 246–253.
[33] Haibin Ling and Kazunori Okada. 2007. An efficient Earth Mover's Distance algorithm for robust histogram comparison. *IEEE Pattern Analysis and Machine Intelligence* 29, 5 (2007), 840–853.
[34] Vebjorn Ljosa, Arnab Bhattacharya, and Ambuj K Singh. 2006. Indexing spatially sensitive distance measures using multi-resolution lower bounds. *Proc. of International Conference on Extending Database Technology (EDBT)* (2006).
[35] Kathy Macropol, Petko Bogdanov, Ambuj K Singh, Linda Petzold, and Xifeng Yan. 2013. I act, therefore I judge: Network sentiment dynamics based on user activity change. *Proc. of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2013), 396–402.
[36] Kathy Macropol, Petko Bogdanov, Ambuj K Singh, Linda Petzold, and Xifeng Yan. 2013. I act, therefore I judge: Network sentiment dynamics based on user activity change (Supplemental Material). http://cs.ucsb.edu/~dbl/papers/sentimentappendix.pdf. (2013). accessed: 2018-05-28.
[37] Andrew McGregor and Daniel Stubbs. 2013. Sketching Earth Mover's Distance on graph metrics. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, M. Goemans, K. Jansen, J.D.P. Rolim, and L. Trevisan (Eds.). Springer, 274–286.
[38] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. 2002. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *Proc. of IEEE International Conference on Data Engineering (ICDE)*. 117–128.
[39] Russell Merris. 1994. Laplacian matrices of graphs: A survey. *Linear Algebra and Its Applications* 197 (1994), 143–176.
[40] Anis Najar, Ludovic Denoyer, and Patrick Gallinari. 2012. Predicting information diffusion on social networks with partial knowledge. In *Proc. of The Web Conference Companion (WWW Companion)*, Leslie Car (Ed.). 1197–1204.
[41] Ofir Pele and Michael Werman. 2008. A linear time histogram metric for improved SIFT matching. In *Proc. of European Conference on Computer Vision (ECCV)*. Springer, 495–508.
[42] Ofir Pele and Michael Werman. 2009. Fast and robust Earth Mover's Distances. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. 460–467.
[43] Anton V. Proskurnikov and Roberto Tempo. 2017. A tutorial on modeling and analysis of dynamic social networks. Part I. *Annual Reviews in Control* 43 (2017), 65 – 79.
[44] Anton V. Proskurnikov and Roberto Tempo. 2018. A tutorial on modeling and analysis of dynamic social networks. Part II. *Annual Reviews in Control* 45 (2018), 166 – 190.

[45] Stephen Ranshous, Shitian Shen, Danai Koutra, Steve Harenberg, Christos Faloutsos, and Nagiza F Samatova. 2015. Anomaly detection in dynamic networks: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics* 7, 3 (2015), 223–247.

[46] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision* 40, 2 (2000), 99–121.

[47] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda. 2011. Detecting changes in opinion value distribution for voter model. In *Proc. of Social Computing, Behavioral-Cultural Modeling and Prediction*. 89–96.

[48] Santiago Segarra, Weiyu Huang, and Alejandro Ribeiro. 2015. Diffusion and superposition distances for signals supported on networks. *IEEE Signal and Information Processing over Networks* 1, 1 (2015), 20–32.

[49] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proc. of ACM Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*. 1397–1405.

[50] Yu Tang, U Leong Hou, Yilun Cai, Nikos Mamoulis, and Reynold Cheng. 2013. Earth Mover's Distance based similarity search at scale. *The VLDB Journal* 7, 4 (2013), 313–324.

[51] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. In *Proc. of ACM Conference on Information and Knowledge Management (CIKM)*. 1031–1040.

[52] Richard C Wilson, Edwin R Hancock, and Bin Luo. 2005. Pattern vectors from algebraic graph theory. *IEEE Pattern Analysis and Machine Intelligence* 27, 7 (2005), 1112–1124.

[53] Richard C. Wilson and Ping Zhu. 2008. A study of graph spectra for comparing graphs and trees. *Pattern Recognition* 41, 9 (2008), 2833 – 2841.

[54] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y Zhao, and Yafei Dai. 2014. Uncovering social network sybils in the wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8, 1 (2014), 2.

Received May, 2018; revised Nov, 2018; revised May, 2019; accepted May, 2019