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

Victor Amelkin* University of California, Santa Barbara

Petko Bogdanov† University at Albany—SUNY

Ambuj K. Singh* University of California, Santa Barbara

***Abstract*—Modeling and predicting people's opinions plays an important role in today's life. For viral marketing and political strategy design, it is particularly important to be able to analyze competing 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 "behaved"' abnormally? Furthermore, can we predict how the opinions of individual users will change in the future?* To answer such questions, it is insufficient to study individual user behavior, since opinions spread beyond users' ego-networks. Instead, we need to consider the opinion dynamics of all users simultaneously. 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 opinion dynamics model. SND has a rich semantics of a transportation problem, yet, is computable in pseudo-linear time, thereby, being applicable to large-scale social networks analysis. We demonstrate the effectiveness of SND in experiments with Twitter data.**

I. Introduction

Modeling and predicting people's opinions plays an important role in today's life. For applications in marketing and political strategy design, it is particularly important to be able to analyze competing 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 "behaved"' abnormally? Furthermore, can we predict how the opinions of individual users will change in the future?* To answer such questions, we need a distance measure for the comparison of states of a social network that would model user opinion evolution taking into account both the location of user opinions as well as the pathways for their likely dissemination. In this work, we develop such a distance measure and employ it for anomaly detection and opinion prediction in Twitter data.

While the dynamics of a social network can be characterized by the evolution of both the network's structure and the user opinions, here, we focus on the latter. We posit there are two *polar opinions* in the network, *positive* "+" and *negative* "−". Users having no or an unknown opinion are *neutral*, while those expressing opinion—*active*. The positive, neutral, and negative opinions are quantified as $+1$, 0, and -1, respectively. 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—political parties or smartphone brands—are inherently competing.

Having observed the behavior of social network users over time and quantified their opinions, we obtain a time series of network states. To analyze it, we 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 show its utility in applications.

To assess the distance between network states, SND takes into account how opinions can propagate in the network. A change of a user's opinion from, say, neutral to positive, contributes to the overall distance between the corresponding network states by reflecting the likelihood of this user's opinion change based on the opinions and locations of other users in the network under a chosen opinion dynamics model. However, since the users interact, the distance measure ought to consider opinion shifts of all users simultaneously. Thus, we design SND as a transportation problem that models opinion propagation in the network. In particular, by making the transportation costs dependent on both the network's structure and the opinions of the users conducting information in the network, we capture the competitive aspect of polar opinion propagation. The summary of *our contributions* is as follows:

▷ We propose SND—the first distance measure suitable for the comparison of social network states containing competing 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 at anomaly detection and user opinion prediction in Twitter data.

II. Network State Comparison and Earth Mover's Distance (EMD)

We propose to address the problem of comparing states of a social network as a transportation problem. The two network states under comparison define supplies and demands, and the costs of opinion transportation are defined based on the shortest paths between the users in the network.

This naturally leads us to one of the well-studied metrics—Earth Mover's Distance (EMD). Originally, defined as a dissimilarity measure for image histograms [9], EMD can be used for the comparison of network states viewed as histograms—maps from the set of network users to the set of possible user opinion values. Intuitively, EMD measures the costs of optimal transformation of one histogram into another with respect to the *ground distance* specifying the costs of moving mass between bins. In our case, the ground distance will be defined based on the shortest paths between the users of the network, computed based on the network's structure and the opinions of the users facilitating opinion propagation.

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

© 2017 IEEE ICDE (https://doi.org/10.1109/ICDE.2017.64)

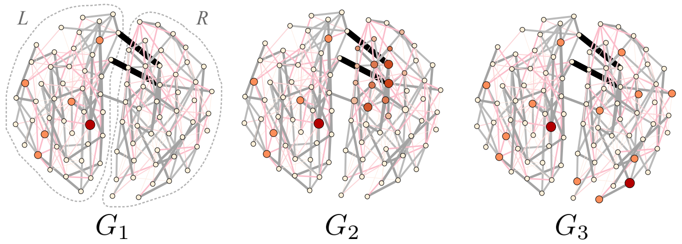


Fig. 1. Three histograms defined over the same two-cluster network.

Formally, given two histograms $P \in \mathbb{R}^{+^n}$ and $Q \in \mathbb{R}^{+^m}$, and ground distance $D \in \mathbb{R}^{+^{n\times m}}$, EMD is the solution to the problem of optimal mass transportation from suppliers $\{P_i\}$ to consumers $\{Q_j\}$ w.r.t. transportation costs $\{D_{ij}\}$:

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

$$\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}D_{ij} \to \min,\quad \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = \min\Big\{\sum_{i=1}^{n} P_i, \sum_{j=1}^{m} Q_j\Big\},$$

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

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

Despite the appeal of EMD, it is limited in that it cannot adequately compare histograms having different total masses—it ignores the mass mismatch, treating a histogram with a very small mass and *any* other histogram as almost identical. The problem is particularly pronounced for social networks, where different states of a social network usually have different total mass due to neutral users' acquiring positive or negative opinions. In our full paper [3], we propose the Generalized Earth Mover's Distance (EMD*), that adequately compares network states with different total mass. In particular, it improves upon existing versions of EMD in that its histogram mass mismatch penalty depends not only on the number of newly activated users, but also on where these users reside in the networks. The latter is achieved through "spreading" the mass mismatch over the network, thereby, incorporating it into the structure of the transportation problem.

To illustrate the difference of EMD* from existing versions of EMD, consider the example in Fig. 1. There are three histograms 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 histograms, 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 process of mass propagation, then G_2 should intuitively be closer to G_1 than G_3. However, only EMD* captures this intuition in that EMD$^\star(G_1, G_2) <$ EMD$^\star(G_1, G_3)$, while for other EMDs, G_2 and G_3 are either equidistant from or identical to G_1.

In the following section, we use EMD* to construct our distance measure for network states containing polar opinions.

III. Social Network Distance (SND)

Given a network $G = \langle V, E\rangle$, where V ($|V| = n$) is the set of nodes (users) and E is the set of edges (social ties), we want to compute the distance between two of its states $P = [P_1, \ldots, P_n]$ and $Q = [Q_1, \ldots, Q_n]$, where P_i and Q_i are the opinions of the i'th user in states P and Q, respectively. Prior to defining SND, we will make two *assumptions*:

▷ *For a given pair of network states, the costs (or likelihood) of opinion propagation depend only on the opinions of the currently active users, taking no account of the potential change of user opinions in the process of opinion transportation.* This assumption, reasonable when the two network states are not very far apart in time, allows us to use the transportation problem as our model, since the transportation costs are required to be static.

▷ *We assume that the users adopting, say, opinion "+" have been affected only by others having the same opinion in the process of opinion propagation. Similarly, suppliers only propagate opinions to the consumers of the same type.* This assumption allows to ask a question about the most likely opinion propagation scenario for positive and negative opinions separately, solving two independent transportation problems, that share the same ground distance (which depends on opinions of both types, capturing opinion competition).

The first assumption allows us to define the ground distance. The cost of opinion propagation from user u to user v depends on their topological proximity, how frequently they communicate, persuasiveness, and stubbornness of u and v as well as the users "separating" them. Formally, the ground distance $D(G_i, op) \in \mathbb{R}^{+^{n\times n}}$, reflecting the costs of propagating opinion op through a network in state G_i, is a matrix containing the lengths of the shortest paths computed in a network with adjacency matrix

$$A^{ext}(G_i, op) = \\ -\log \mathbb{P}(G_i, op) - \log \mathbb{P}^{in}(G_i, op) - \log \mathbb{P}^{out}(G_i, op), \tag{2}$$

where the summands on the right are n-by-n matrices of log-probabilities of *communication*, *opinion adoption*, and *opinion spreading*, respectively. Communication probabilities $\mathbb{P}(G_i, op)$ can be defined as the relative frequencies of communication between users, provided that it is known how often they actually interact. Opinion adoption probabilities $\mathbb{P}^{in}(G_i, op)$ reflect users' susceptibility/stubbornness. The simplest way to define the opinion spreading penalties $-\log \mathbb{P}^{out}(G_i, op)$ is in a model-agnostic fashion as follows.

$$-\log \mathbb{P}^{out}_{uv}(G_i, op) = \begin{cases} c_{adverse} & \text{if } G_i[u] = -op \vee G_i[v] = -op, \\ c_{neutral} & \text{if } G_i[u] = 0, \\ c_{friendly} & \text{if } G_i[u] = op \wedge G_i[v] \neq -op, \end{cases}$$

where $c_{adverse}, c_{neutral}, c_{friendly} \in \mathbb{R}^+$ are constant penalties for spreading opinion op by the users having, respectively, adverse, neutral, or friendly opinion relatively to op, and $G_i[u]$ is the opinion of user u in network state G_i. This simple definition implies that users willingly spread opinions similar to their own ($c_{friendly}$ is small); are unwilling to spread adverse opinions ($c_{adverse}$ is large); with neutral users' behavior being somewhere in-between ($c_{friendly} < c_{neutral} < c_{adverse}$). Alternatively, $\mathbb{P}^{out}(G_i, op)$ can be defined based on any existing opinion dynamics model. In our full paper [3], we provide such cost definitions based on the version of the Independent Cascade [5] and the Linear Threshold [4] models allowing for competing opinions.

Finally, we can formally define SND.

$$\boxed{\begin{gathered}\mathrm{SND}(G_1, G_2) = \tfrac{1}{2}\times \\ \big[\mathrm{EMD}^\star(G_1^+, G_2^+, D(G_1, +)) + \mathrm{EMD}^\star(G_1^-, G_2^-, D(G_1, -)) + \\ \mathrm{EMD}^\star(G_2^+, G_1^+, D(G_2, +)) + \mathrm{EMD}^\star(G_2^-, G_1^-, D(G_2, -))\big],\end{gathered}} \quad (3)$$

where the users holding negative opinions are considered neutral in G_i^+, and the users holding positive opinions are neutral in G_i^-. Notice that SND depends on both ground distances $D(G_1, op)$ and $D(G_2, op)$. One reason for it is that network states G_1 and G_2 may be time-unordered, and it may be unknown which of two network states corresponds to the past and defines the likelihoods of different opinion propagation scenarios. Another reason for such a choice is SND's symmetry w.r.t. G_1 and G_2, which makes SND metric.

IV. Efficient Computation of SND

SND is defined (3) as a linear combination of several instances of EMD*, and, thus, computation of SND involves:

▷ Computing the ground distance $D(G_i, op)$ 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_i.

▷ Computing EMD*, when the network states and the ground distance are provided.

Computing the ground distance D implies computing shortest paths, whose direct computation for all pairs of users using Dijkstra's algorithm would incur time cost $\mathcal{O}(n^2 \log n)$ 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, or a solver that exploits the special structure of the transportation problem, such as the transportation simplex algorithm. The complexity of both algorithms is, however, 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 histograms derived from a social network's states, since they either drastically simplify the ground distance, or are effective only for some graphs, such as trees, structurally not characteristic of social networks. Nevertheless, in what follows, we propose a method to exactly compute SND 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 G_1 and G_2 under comparison is significantly smaller than the total number n of users in the network. This assumption is reasonable, because in most applications the network states under comparison are not very far apart in time and, hence, $n_\Delta \ll n$.

Assumption 2: The opinion transportation costs, defined as the elements of adjacency matrix A^{ext} in (2), are positive integers bounded from above by constant $U \ll +\infty \in \mathbb{Z}^+$. This assumption is easy to satisfy by the appropriate choice of costs, and does not limit our analysis.

As computing SND is equivalent to computing four instances of EMD*, our focus here is on efficient computation of EMD* on the inputs supplied by SND. Our method for computing SND—given as Theorem 1 below—requires the following lemma. The full proofs are provided in [3].

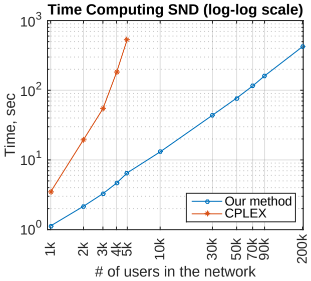


Fig. 2. Number n_Δ of users holding different opinion is fixed at 1000; the network size n grows up to 200k.

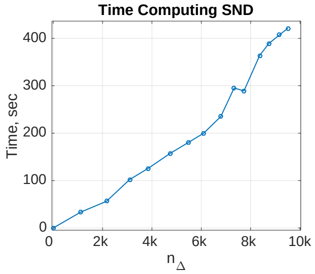


Fig. 3. The network size n is fixed at 20k; the number n_Δ of users holding different opinions grows up to 10k.

Lemma 1. *Given two arbitrary histograms $P, Q \in \mathbb{R}^n$ and a ground distance $D \in \mathbb{R}^{n\times n}$, if D is semimetric (a metric with symmetry requirement dropped), then for any $i \in \{1, \ldots, n\}$, the following holds*

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

Theorem 1. *Under Assumptions 1 and 2, SND between network states $P = [P_1, \ldots, P_n]$ and $Q = [Q_1, \ldots, Q_n]$ defined over network $G = \langle V, E\rangle$, ($|V| = n, |E| = m$) can be exactly computed in time $\mathcal{O}(n_\Delta(m + n\sqrt{\log U} + n_\Delta^2 \log(n_\Delta nU)))$. In a sparse network, with bounded n_Δ, this time is $\mathcal{O}(n)$.*

Proof Idea: SND has EMD* at its core, and here we are concerned with the efficient computation of the latter. To efficiently compute, say, EMD$^\star(P^+, Q^+, D(P, +))$, we, first, apply Lemma 1 to P^+ and Q^+, whose effect is changing the values of many bins in P^+ and Q^+ to zeros, without affecting the value of EMD* between them. The latter is beneficial, since zero bins are discarded from the underlying transportation problem. As a result only one of P^+ and Q^+ still has the number $\mathcal{O}(n)$ of bins proportional to the number of users in the network, while the other one's number $\mathcal{O}(n_\Delta)$ of bins is proportional to the number of users whose opinions are different in P^+ and Q^+ (and, due to Assumption 1, $n_\Delta \ll n$). Thus, the computation of $D(P, +)$ corresponds to n_Δ single-source shortest path computations, which, due to Assumption 2, can be performed using Dijkstra algorithm using a combination of a radix and Fibonacci heaps [1], with time complexity $T_{sssp} = \mathcal{O}(m + n\log\sqrt{U})$; and the computation of EMD$^\star(P^+, Q^+, D(P, +))$ accounts for solving an unbalanced ($n_\Delta \ll n$) transportation problem using Goldberg-Tarjan's min-cost flow algorithm [7] augmented with the two-edge push rule of Ahuja et al. [2] in time $T_{transp} = \mathcal{O}(n_\Delta m + n_\Delta^3 \log(n_\Delta \max_{i,j} D(P, +)_{ij}))$. Thus, the total time for computing EMD* and, hence, SND is $T = \mathcal{O}(n_\Delta T_{sssp} + T_{transp}) = \mathcal{O}(n_\Delta(m + n\log\sqrt{U} + n_\Delta^2 \log(n_\Delta nU)))$. Observing that in a sparse network $m = \mathcal{O}(n)$ concludes the proof. ■

We have implemented SND in MATLAB and C++ (available at http://cs.ucsb.edu/~victor/pub/ucsb/dbl/snd/). We use the min-cost network flow solver CS2 [6] that implements Goldberg-Tarjan's algorithm [7], but, unlike it is prescribed by Theorem 1, does not use the two-edge push rule [2]. As a result, our implementation of SND scales slightly worse than linearly as guaranteed by Theorem 1, but still very well to be applicable to real-world social networks. Fig. 2 shows how our implementation of SND based on Theorem 1 scales in the number n of users in the network in comparison with a direct computation of SND using CPLEX' linear solver. Our implementation's scalability in the number n_Δ of users who have changed their opinion is shown in Fig. 3.

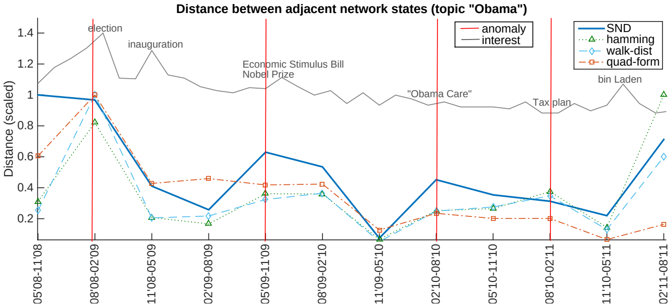
V. Experimental Results

A. *Detecting Anomalous Network States*

In a series of network states $G_1, G_2, \dots, G_n$, we want to detect which network state transitions $\langle G_i, G_{i+1} \rangle$ are anomalous in the sense of not following the expected opinion evolution. To this end, we compute the distances between adjacent network states, normalize these distances by the number of active users, and rescale to fit $[0;1]$. Then, spikes in the resulting series of distances are considered anomalies.

We analyze a subset of tweets from Twitter dataset [8] sent between May-2008 and August-2011, containing hashtag "Obama", and connect users in a network based on their follower-followee relationship. As a result, we obtain a network of 10k users, each having an average of 130 neighbors. Within each quarter, we quantify the sentiment of each tweet; subsequently, the opinions of all the users comprise that quarter's network state. The "search interest" data from Google Trends was used as the ground truth.

We compare the performance of SND at anomaly detection with that of several other distance measures: Hamming distance, quadratic-form distance, and walk-distance (the average amount by which the opinion of a user differs from the average opinion of its active neighbor). The anomaly detection results are shown in the following Figure.



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, 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 still remains a very controversial topic.

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". (a) The 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, so, as expected, SND, while indicating it as an anomaly, does not perform any better at detecting that event than the simpler distance measures sensitive to the user activation rate. The undetected by SND (b) Obama's tax cut extension and (c) bin Laden's death, however, were not polarizing—the tax cut had received large support in the Senate from both the Democrats and the Republicans, while bin Laden's death has unlikely been perceived differently by the US users of Twitter.

B. *Predicting User Opinions*

Given a series of states of a social network, we want to predict the unknown opinions of individual users in the current network state G_0 based on the observed recent G_{-t} ($t \in \mathbb{N}$) and the (incomplete) current network states. *We assume that during the periods corresponding to the observed recent network states G_{-t}, the network evolved "smoothly"*, so the recent past network states are informative of the current network state. Under this assumption, we compute distances $dist(G_{-t}, G_{-t+1})$ between adjacent past network states, then, extrapolate the obtained series of distances to estimate the distance d^* from the most recent G_{-1} to the yet unknown *complete* current network state. Then, we search for the assignment of opinions to the target users in the current network state that would make the distance $dist(G_{-1}, G_0^*)$ from the most recent to the modified current network state as close to estimate d^* as possible. In each experiment, we uniformly randomly select 20 active users—with equal representation of positive and negative opinions—in the current network state, predict their opinions and measure the prediction accuracy. This procedure is repeated for different sets of active users, and mean accuracies and standard deviations are reported. The predictions are made using SND as well as other distance measures. The opinion prediction results are summarized below.

Distance Measure	Opinion Prediction Accuracy, %	
	mean	std
SND	**75.63**	**5.60**
hamming	68.13	5.80
quad-form	67.50	9.63
walk-dist	31.88	9.98

References

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

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

[3] V. Amelkin, A. K. Singh, and P. Bogdanov, "A distance measure for the analysis of polar opinion dynamics in social networks (Full Paper)," *available at* http://cs.ucsb.edu/~victor/pub/ucsb/dbl/snd/snd-full.html.

[4] A. Borodin, Y. Filmus, and J. Oren, "Threshold models for competitive influence in social networks," in *Proc. International Workshop on Internet and Network Economics*. Springer, 2010, pp. 539–550.

[5] T. Carnes, C. Nagarajan, S. M. Wild, and A. Van Zuylen, "Maximizing influence in a competitive social network: a follower's perspective," in *Proc. ACM Electronic Commerce*, 2007, pp. 351–360.

[6] A. Goldberg, "An efficient implementation of a scaling minimum-cost flow algorithm," *Journal of Algorithms*, vol. 22, no. 1, pp. 1–29, 1997.

[7] A. Goldberg and R. Tarjan, "Solving minimum-cost flow problems by successive approximation," in *Proc. ACM STOC*, 1987, pp. 7–18.

[8] K. Macropol, P. Bogdanov, A. K. Singh, L. Petzold, and X. Yan, "I act, therefore I judge: Network sentiment dynamics based on user activity change," *Proc. ACM ASONAM*, pp. 396–402, 2013.

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