ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746556

EMBEDDING SIGNALS ON GRAPHS WITH UNBALANCED DIFFUSION EARTH MOVER'S DISTANCE

Alexander Tong[1,*] *Guillaume Huguet*[2,*] *Dennis Shung*[3,*] *Amine Natik*[2] *Manik Kuchroo*[4]
Guillaume Lajoie[2] *Guy Wolf*[2,†] *Smita Krishnaswamy*[4,1,†]

[1] Yale University, Dept. of Comp. Sci.; [3] Dept. of Medicine; [4] Dept. of Genetics, New Haven, CT, USA
[2] Université de Montréal, Dept. of Math. & Stat.; Mila – Quebec AI Institute, Montreal, QC, Canada

ABSTRACT

In modern relational machine learning it is common to encounter large graphs that arise via interactions or similarities between observations in many domains. Further, in many cases the target entities for analysis are actually signals on such graphs. We propose to compare and organize such datasets of graph signals by using an earth mover's distance (EMD) with a geodesic cost over the underlying graph. Typically, EMD is computed by optimizing over the cost of transporting one probability distribution to another over an underlying metric space. However, this is inefficient when computing the EMD between many signals. Here, we propose an unbalanced graph EMD that efficiently embeds the unbalanced EMD on an underlying graph into an L^1 space, whose metric we call *unbalanced diffusion earth mover's distance (UDEMD)*. Next, we show how this gives distances between graph signals that are robust to noise. Finally, we apply this to organizing patients based on clinical notes, embedding cells modeled as signals on a gene graph, and organizing genes modeled as signals over a large cell graph. In each case, we show that UDEMD-based embeddings find accurate distances that are highly efficient compared to other methods.

***Index Terms*—** Optimal transport, graph signal processing, knowledge graph, graph diffusion

1. INTRODUCTION

The task of comparing probability distributions is applicable to a wide variety of machine learning problems, giving rise to popular ϕ-divergences such as the Kullback-Leibler (KL), Hellinger, or total variation (TV) divergences, which ignore the underlying geometry of their support. The Earth Mover's Distance (EMD), also known as the Monge-Kantorovich or Wasserstein Distance, explicitly takes into account this underlying geometry via a domain-specific ground distance, which has many advantages on empirical probability distributions [1, 2]. Here, we show that earth mover's distances are useful in a new domain: that of graph signals. In modern relational machine learning, we encounter large graphs that arise via interactions between entities in many domains [3, 4]. Features of such entities can be considered as signals on the graph. For such signals, which often tend to be noisy, we propose a new unbalanced graph earth mover's distance, and use it to organize the signals and determine relationships between them.

Since graphs can contain tens (Cora) [3] to hundreds of thousands of nodes (SNOMED-CT) [4], there is a great need for this measure to be computationally efficient. While the Wasserstein distance is intuitively attractive, it presents computational challenges. Here, based on the recent diffusion EMD method [5], we show that an efficient unbalanced EMD between signals can be computed as the difference between graph convolutions of the signal with multiscale graph kernels. This unbalanced EMD can be computed in linear time with convergence guarantees and without solving an optimization problem. We call our distance *unbalanced diffusion earth mover's distance* (UDEMD).

While previous work on Wasserstein distance embedding mostly focused on its relation to the balanced optimal transport problem [5–9], we propose an unbalanced Wasserstein embedding approach between large number of distributions defined as signals on graphs. Since graph signals tend to be noisy, an *unbalanced* transport, which can choose not to transport parts of the data space when it is inefficient to do so, leads to more robust distances between graph distributions that are less sensitive to outliers in the signal.

We apply UDEMD to medical knowledge graphs using Systemized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [4]. We show that unbalanced diffusion EMD can be used to find meaningful distances between patients which successfully clusters patients into different diagnosis categories, and allows us to find relationships between patient features. We also apply UDEMD to single cell RNA-sequencing data where we can model both cells as signals on gene interaction graphs or genes as signals on cell-similarity graphs. In cases where the gene regulatory network is well known, researchers have shown that affinity between cells can be computed as an earth mover's distance [10, 11]. We show that UDEMD runs orders of magnitude faster than the Sinkhorn and network simplex methods used in those works, while maintaining accuracy. In cases where the gene regulatory network is not well known, we model the transposed problem, deriving groupings of genes that function similarly by modeling genes as expression values over single cells. Here, we show that the UDEMD provides robust distances that recapitulate ground truth gene groupings in single cell data from peripheral blood mononuclear cells (PBMCs).

2. PRELIMINARIES

In this section we review the Wasserstein metric, embedding based methods for approximating it, and unbalanced optimal transport.

The Wasserstein metric is a notion of distance between two mea-

[*] Equal contribution; [†] Equal senior author contribution. This research was partially funded by NIH grant K23DK125718-01A1 [*D.S.*]; IVADO PhD Excellence Scholarship [*A.N.*]; CIFAR AI Chair, NSERC Discovery grant 03267 [*G.W.*]; Chan-Zuckerberg Initiative grants 182702 & CZF2019-002440 [*S.K.*]; NSF career grant 2047856 [*S.K.*]; Sloan Fellowship FG-2021-15883 [*S.K.*]; and NIH grants R01GM135929 & R01GM130847 [*G.W., S.K.*]. The content provided here is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. Correspondence to <guy.wolf@umontreal.ca> and <smita.krishnaswamy@yale.edu>

sures μ, ν on a measurable space Ω endowed with a metric $d(\cdot,\cdot)$ known as the ground distance. The primal formulation of the Wasserstein distance W_d, also known as the earth mover's distance, is defined as:

$$W_d(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} d(x,y)\pi(dx,dy), \qquad (1)$$

where $\Pi(\mu,\nu)$ is the set of joint probability distributions π on the space $\Omega\times\Omega$, such that for any subset $\omega\subset\Omega$, $\pi(\omega\times\Omega)=\mu(\omega)$ and $\pi(\Omega\times\omega)=\nu(\omega)$. Also of interest is the entropy regularized Wasserstein distance [12], which reduces the computation to $O(n^2)$. This algorithm is extremely parallelizable, and works quite well even for a small number of iterations [2], and there are many works investigating how to scale this to larger problems.

However, when comparing a large number of signals (say m), we must solve the optimization for each pair of signals, i.e. $O(m^2)$ optimizations. For this reason, we turn to methods that approximate the dual of the Wasserstein metric, also known as the Kantorovich-Rubenstein dual formulation, which relies on witness functions. Many works optimize the cost over a modified family of witness functions such as functions parameterized by neural networks [13–15], functions defined over trees [6, 9], and wavelet bases [7, 8]. An efficient algorithm recently proposed is Diffusion EMD [5], it is based on a multi-scale representation of the signals. Indeed, it can be seen as a weighted average of the L^1 distances between two signals at different scales.

There are numerous formulations of *unbalanced* optimal transport both to accommodate problems with unequal masses and to provide robustness to outlier points [1, 16]. In general these can be formulated as a mixture between a pure optimal transport problem and a ϕ-divergence. We focus on the formulation using the total variation, referred to as the TV-unbalanced problem:

$$\text{TV-UW}_d(\mu,\nu) = \inf_s \left\{ \mathrm{W}_d(\mu+s,\nu) + \lambda \mathrm{TV}(\mu+s,\mu) \right\}, \qquad (2)$$

where $\lambda = \min\{\lambda_\mu, \lambda_\nu\}$ and λ_μ, λ_ν control the relative cost of mass creation / destruction compared to transportation. Intuitively, we can think of Eq. 2 as minimizing over the "teleporting" mass s, that is too costly to transport.

In the unbalanced optimal transport literature, most often considered is the KL-divergence formulation which can be solved efficiently in the case of entropic regularized problem [17–19], but is difficult to optimize stochastically as is possible in the balanced case, limiting scalability [20, 21]. The TV-unbalanced problem (Eq. 2) can be solved by adding a "dummy point" that is connected to every point with equal cost [22, 23]. However, adding a dummy point removes the metric structure necessary for dual-based Wasserstein distances. It is not immediately obvious that Eq. 2 is efficiently computable while maintaining this structure. To address this issue, [24] showed that the TV-unbalanced problem can be solved through cost truncation. Following their work, we will show that there is an embedding of distributions to vectors where the L^1 distance between vectors is equivalent to the TV-UW between the distributions.

3. UNBALANCED DIFFUSION EARTH MOVER'S DISTANCE

While Diffusion EMD presented in [5] can provide an earth mover's distance between graph signals, its formulation is not motivated by considering noisy signals on graphs or outliers, but rather geared to avoid high dimensional density estimation. Here, we focus on utilizing EMD to organize graph signals. Therefore, we are interested in distances that are immune to outlier spikes in the signals. While the multiscale smoothing proposed in [5] is effective in handling noisy perturbation of the signals, it is less effective at dealing with outlier vertex components of the signal. However, as we show here, the construction can be adapted to consider unbalanced transport, which is essentially based on the idea that a more faithful earth mover's distance is given by a transport in which we ignore some of the mass – particularly, mass that requires large transport costs. To incorporate this idea, we modify the formulation of [5] by only considering certain scales. This yields the Unbalanced Diffusion EMD (UDEMD), which is topologically equivalent to the total variation unbalanced Wasserstein distance.

Definition 1. *The Unbalanced Diffusion Earth Mover's Distance (UDEMD) between two signals μ, ν is*

$$\text{UDEMD}_{\alpha,K}(\mu,\nu) := \sum_{v\in V}\sum_{k=0}^{K} \|g_{\alpha,k}(\mu(v)) - g_{\alpha,k}(\nu(v))\|_1 \quad (3)$$

where $0<\alpha<1/2$ is a meta-parameter used to balance long- and short-range distances, and

$$g_{\alpha,k}(\mu(v)) := 2^{-(K-k-1)\alpha}\left(\boldsymbol{\mu}^{(k+1)} - \boldsymbol{\mu}^{(k)}\right) \qquad (4)$$

where $\boldsymbol{\mu}^{(t)}$ is short for $\boldsymbol{\mu}^{(t)}(v) = (\boldsymbol{P}^{2^t}\mu)(v)$ and K is the maximum scale considered.

The scale K relates to the unbalancing threshold (see Fig. 1 and discussion in Sec. 3.1). In practice, α is set close to 1/2, hence we drop the subscript and use the notation UDEMD_K.

3.1. Equivalence to (unbalanced) Wasserstein distance

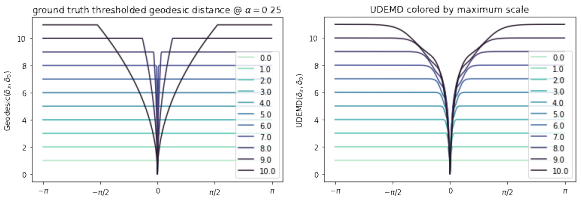
In [24], it was shown that truncated-cost optimal transport distances were equivalent to unbalanced Wasserstein distances, and that they are useful in outlier detection. However, there the proposed implementation used a truncated matrix with the standard Sinkhorn algorithm [12]. Here we show a similar result for the Unbalanced Diffusion EMD from Def. 1, i.e., showing that with scale truncation it is equivalent to an unbalanced Wasserstein distance. We first adapt Theorem 3.1 from [24] in the following Lemma 1, which will in turn be combined with Lemma 2 to yield this result.

Lemma 1. *The Wasserstein distance with a truncated ground distance $d_\lambda(x,y) = \min\{\lambda, d(x,y)\}$ for some constant λ and distance d is equivalent to a total variation unbalanced Wasserstein distance for some constant λ, i.e., $W_{d_\lambda}(\mu,\nu) = \text{TV-UW}_d(\mu,\nu)$.*

The theory developed in [5] assumed that the support of the considered distributions was a closed Riemannian manifold. In such a case, Diffusion EMD will converge to a distance that is equivalent to the Wasserstein distance defined with the geodesic on the manifold. The following Lemma extends this theory to show that UDEMD (Def. 1) will converge to a Wasserstein distance where the ground distance is a thresholded geodesic.

Lemma 2. $\text{UDEMD}_K(\mu,\nu)$ *approximates a metric equivalent to the Wasserstein distance $W_{d_\lambda}(\mu,\nu)$, defined as in Lemma 1, with the ground distance being a truncated geodesic distance on the manifold, i.e., $d_\lambda(x,y) = \min\{\lambda, \rho(x,y)\}$ for $\lambda > 0$.*

Proof. We present a proof sketch here; the main part of the proof follows the same lines as in Corollary 3.1 of [5]. In Def. 1, an anisotropic kernel $\boldsymbol{P}$ is used, which can be shown to converge to the



(a) Thresholded geodesic (b) UDEMD between diracs

Fig. 1: On a ring graph $n = 500$ compares the UDEMD to the thresholded ground distance, this suggests that UDEMD closely approximates the thresholded ground distance with $\lambda \approx 2^K$.

Heat kernel on a Riemannian manifold (see [25], Prop. 3). In [26], it is shown that the construction of Def. 1 using the Heat Kernel will converge to a metric that is equivalent to the Wasserstein with ground distance $\min\{1, \rho(x,y)^{2\alpha}\}$, where ρ is the geodesic on the manifold. Because the metrics $\min\{1, \rho(x,y)^{2\alpha}\}$ and $\min\{\lambda, \rho(x,y)^{2\alpha}\}$ are equivalent for $\lambda > 0$, the Wasserstein distances induced by these metrics are also equivalent. □

By combining Lemmas 1 and 2, we have that the UDEMD from Def. 1 approximates a metric equivalent to an unbalanced optimal transport metric. Formally, using the equivalence notation from [5], we have $\text{UDEMD}_K(\mu,\nu) \simeq \text{TV-UW}_d(\mu,\nu)$. We note that while our result here establishes a relation between these two metrics, it does not directly quantify the relation between the λ and K. We leave careful theoretical and rigorous study of this relation to future work[1], but mention here that we observe empirically, as shown in Fig. 1, the choice of K indeed acts in a similar way to the threshold λ on the ground distance.

Algorithm 1 $\text{UDEMD}(\boldsymbol{A}, \boldsymbol{\mu}, K, \alpha) \to \boldsymbol{b}$

Input: $n \times n$ graph adjacency $\boldsymbol{A}$, $n \times m$ distributions $\boldsymbol{\mu}$, maximum scale K, and snowflake constant α.
Output: $m \times (K+1)n$ distribution embeddings $\boldsymbol{b}$
$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$
$\boldsymbol{\mu}^{(2^0)} \leftarrow \boldsymbol{\mu}$
for $k = 1$ **to** K **do**
  $\boldsymbol{\mu}^{(2^k)} \leftarrow \boldsymbol{P}^{2^k}\boldsymbol{\mu}^{(2^{k-1})}$
  $\boldsymbol{b}_{k-1} \leftarrow 2^{(K-k-1)\alpha}(\boldsymbol{\mu}^{(2^k)} - \boldsymbol{\mu}^{(2^{k-1})})$
end for
$\boldsymbol{b}_K \leftarrow \boldsymbol{\mu}^{(2^K)}$
$\boldsymbol{b} \leftarrow [\boldsymbol{b}_0, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_K]$
$\tilde{\boldsymbol{b}} \leftarrow \text{Subsample}(\boldsymbol{b})$

To compute the UDEMD defined in Def. 1, we present Alg. 1 with time complexity $O(2^K m|E|)$, which is similar to algorithms used in graph neural networks. Our algorithm scales well with the size of the graph, the number of distributions m and number of points n, but poorly with the maximum scale K. We note that the maximum scale considered for Diffusion EMD was of order $O(\log|V|)$, derived from the convergence of the heat kernel to its steady state. Here, on the other hand, we decouple the tuning of K and find that a much smaller maximum scale suffices, and in fact (as discussed in Sec. 3.1) corresponds to a well characterized unbalanced earth mover's distance on the underlying geometry of the graph. This leads to Alg. 1 emphasizing preferable scaling properties for small K, and easily accelerated by computation on GPUs.

[1]We deem the required nuanced treatment of scaling constants in equivalence bounds out of scope in this work.

4. RESULTS

In this section, we show that UDEMD is an efficient and robust method for measuring distances between graph signals and then using the distances to find embeddings and organization of the signals (often entities such as patients). We compare UDEMD to a GPU implementation of numerically stabilized Sinkhorn optimization that includes minibatching of sets of distributions. However, despite this, this method runs out of memory when there are beyond 10,000 nodes in the graph. We note that all methods of this type require solving m^2 optimizations, even when looking for nearest neighbors. Unless otherwise noted, we set $K = 4$ and $\alpha = 1/2$.

Spherical data test case To test the speed and robustness of UDEMD we begin with a dataset where we have knowledge of the intrinsic ground distances and can vary the number of points and distributions. For this dataset we sample m Gaussian distributions with means distributed uniformly on the unit sphere with 10 points each for a total of $n = 10m$ points. We add a random noise spike at a uniformly random location on the sphere to check robustness to this type of noise. The goal is to predict the neighboring distributions on the sphere. We find that UDEMD is significantly more scalable and find that there is a sweet spot in terms of K at $K = 4$ for this dataset. The UDEMD with $K = 4$ performs significantly better than the balanced Diffusion EMD case with this type of noise. This supports the claim that setting $K \ll O(\log n)$ is beneficial in real world datasets. UDEMD also outperforms the graph-TV distance as it is both faster and more accurate at $K = 1$, and more accurate overall.

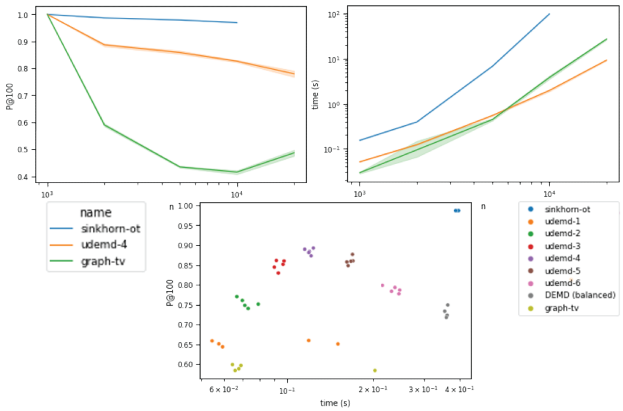
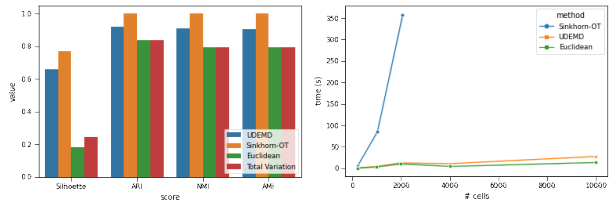


Fig. 2: UDEMD is more scalable than Sinkhorn-OT and performs better than graph total variation. (left) Shows performance as measured by P@100, the fraction of the 100 nearest neighbors predicted correctly, against problem size. (middle) Shows time against problem size, and (right) shows performance vs. time on a problem size of $n = 2000$ for different choices of K.

Single-cell data with cells as signals over gene graphs We consider 206 cells from the K562 human lymphoblast cell line as signals over a known 10000-node gene graph in single-cell RNA seq data [27]. We measure the distance between cells based on their transport on this gene graph. This was recently independently proposed by [10] and [11], who showed that OT over the gene graph can provide better distances between cells than Euclidean measures. We measure the performance of these methods based on how well the resulting distance matrix between cells matches the clusters according to four scores: Silhouette score, the adjusted rand index (ARI), the normalized mutual information (NMI), and the adjusted mutual information (AMI). In Fig. 3, we see that UDEMD performs almost as well as

Sinkhorn-OT, and much better than the Euclidean and total variation distances that do not take into account the gene graph as well as much faster than Sinkhorn-OT, scaling almost as well as Euclidean distances due to the embedding. Note however, that using balanced transport (see Fig. 2 right) degrades the accuracy. The balanced transport compared here is the original Diffusion EMD from [5] which is not a thresholded distance, and thus noise in the data are able to perturb the accuracy of the distances.



(a) Performance metrics (b) Computation comparison

Fig. 3: UDEMD achieves better clustering than Euclidean and total variation (TV) distances, and performs similarly well to Sinkhorn-OT but is much more scalable with similar scalability to Euclidean and TV distances. (a) performance in terms of Silhouette score, ARI, NMI, and AMI (b) computation time vs. problem size.

Single-cell data with genes as signals over cell graphs Next we applied our approach to 4,360 peripheral blood mononuclear cells measured via single cell RNAseq publicly available on the 10X platform. We consider three curated gene sets that are explanatory for this dataset. We compare the distances between genes using UDEMD, Euclidean, total variation and Sinkhorn-OT distances. We can see that the genes canonical for monocytes (orange), T cells (green) and B cells (blue) all appear to be closely positioned to one another and separate between the groups in our embedding in contrast to a Euclidean distance embedding of the genes where the clusters are less clear (Fig. 4a). Visualizing the UDEMD distance between our 46 genes in a heatmap, we can identify the three clusters as dark blocks of low distance (Fig. 4b). This result is quantified in (Fig. 4c), where UDEMD performs the best on 3/4 metrics. Last, we tried to see how diffusion time scale impacted silhouette score, identifying that score maximized at a timescale of $K = 10$ and did not improve with higher scales.

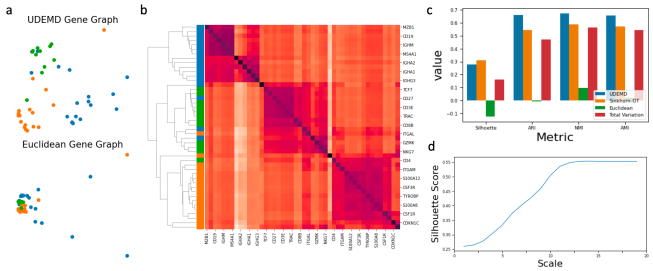


Fig. 4: (a) Visualization of gene graphs of 46 genes canonical for different cell types using UDEMD and Euclidean ground distances (blue for B cells, orange for monocytes and green for T cells), (b) heat map of gene distances (c) clustering performance (d) silhouette score vs. maximum diffusion scale K.

A patient concept knowledge graph We consider a knowledge graph constructed from medical concepts captured in clinical documentation and reporting. SNOMED-CT is a widely used collection of terms and concepts with defined relationships considered to be an international standard for medical concepts captured from the electronic health record. SNOMED-CT has a pre-defined knowledge graph with concept-relation-concept triplets, which we subset to the Clinical Findings concept model (version 3/2021). We used 52,150 discharge summaries from MIMIC-III, which contain all information about a patient's hospital course and extracted concepts using MetaMap (version 2018) [28]. These medical concepts were then used as signals on the SNOMED-CT knowledge graph, which link all relevant concepts together. The metadata used to label patients included primary diagnosis, a physician-designated diagnosis which was stored separately in the MIMIC-III database.

One of the advantages of the UDEMD-based embedding is the identification of clinically meaningful overlaps that may not be apparent from the single primary diagnosis recorded in the database. Patients with a primary diagnosis of intracranial bleeding (bleeding in the brain) can also have primary brain masses and tumors. Compared to the spurious fragmentation of patients with the same diagnosis of intracranial bleeding into several clusters in the TV embedding, UDEMD consolidated patients with the same diagnosis of intracranial bleeding and specifically grouped those that may have had bleeding due to a primary brain mass or tumor (See Fig. 5b). Interestingly, UDEMD also identified patients who were predicted to have intracranial bleeding to have the diagnosis of stroke with higher accuracy, reflecting consistency with the fact that a subtype of stroke (hemorrhagic) is due to intracranial bleeding (Fig. 5d).

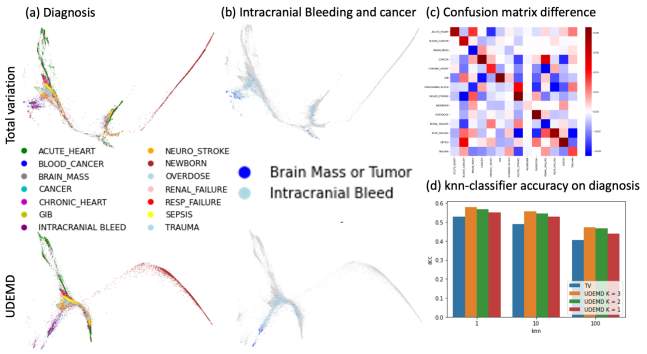


Fig. 5: Embeddings of patients modeled as signals over the SNOMED-CT graph using TV distance (a top) and using UDEMD distance (a bottom), colored by patient diagnosis. UDEMD better organizes the space as noted by selected terms in (b), difference of confusion matrices in (c) and k-nearest neighbors classification accuracy on the diagnosis in (d). In (b) note that the TV embedding (top) creates a spurious separation (due to noise in the signal) between subsets of patients who display intracranial bleeding that is not distinguished by diagnosis. On the other hand the UDEMD embedding (bottom) shows a continuum of patients with this diagnosis. The same holds for patients with brain mass or tumor shown in green.

5. CONCLUSION

In this work we explored the use of earth mover's distance to organize signals on large graphs. We presented an unbalanced extension of Diffusion EMD, which we showed approximates a distance equivalent to the total variation unbalanced Wasserstein distance between signals on a graph. We showed how to compute nearest neighbors in this space in time log-linear in the number of nodes in the graph and the number of signals. Finally, we demonstrated how this can be applied to entities which can be modeled as signals on graphs between genes, cells, and biomedical concepts.

6. REFERENCES

[1] Gabriel Peyré and Marco Cuturi, *Computational Optimal Transport*, arXiv, 2019.

[2] Arturs Backurs, Yihe Dong, Piotr Indyk, Ilya Razenshteyn, and Tal Wagner, "Scalable nearest neighbor search for optimal transport," in *ICML*, 2020.

[3] Aleksandar Bojchevski and Stephan Günnemann, "Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking," in *ICLR*, 2021.

[4] Stefan Schulz and Gunnar O. Klein, "Snomed ct – advances in concept mapping, retrieval, and ontological foundations. selected contributions to the semantic mining conference on snomed ct (smcs 2006)," *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, pp. S1, 2008.

[5] Alexander Tong, Guillaume Huguet, Amine Natik, Kincaid MacDonald, Manik Kuchroo, Ronald Coifman, Guy Wolf, and Smita Krishnaswamy, "Diffusion earth mover's distance and distribution embeddings," in *ICML*, 2021.

[6] Piotr Indyk and Nitin Thaper, "Fast image retrieval via embeddings," in *3rd International Workshop on Statistical and Computational Theories of Vision*, 2003.

[7] Sameer Shirdhonkar and David W. Jacobs, "Approximate earth mover's distance in linear time," in *CVPR*, 2008.

[8] Matan Gavish, Boaz Nadler, and Ronald R Coifman, "Multiscale Wavelets on Trees, Graphs and High Dimensional Data: Theory and Applications to Semi Supervised Learning," in *ICML*, 2010.

[9] Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi, "Tree-Sliced Variants of Wasserstein Distances," in *NeurIPS*, 2019.

[10] Geert-Jan Huizing, Gabriel Peyré, and Laura Cantini, "Optimal Transport improves cell-cell similarity inference in single-cell omics data," *BioRxiv*, 2021.

[11] Riccardo Bellazzi, Andrea Codegoni, Stefano Gualandi, Giovanna Nicora, and Eleonora Vercesi, "The gene mover's distance: Single-cell similarity via optimal transport," *arXiv*, 2021.

[12] Marco Cuturi, "Sinkhorn Distances: Lightspeed Computation of Optimal Transport," in *NeurIPS*, 2013.

[13] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," in *ICML*, 2017.

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville, "Improved Training of Wasserstein GANs," in *NeurIPS*, 2017.

[15] Alexander Tong, Guy Wolf, and Smita Krishnaswamy, "Fixing Bias in Reconstruction-based Anomaly Detection with Lipschitz Discriminators," in *IEEE MLSP*, 2020.

[16] Yogesh Balaji, Rama Chellappa, and Soheil Feizi, "Robust optimal transport with applications in generative modeling and domain adaptation," in *NeurIPS*, 2020.

[17] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, "Unbalanced optimal transport: Dynamic and Kantorovich formulations," *Journal of Functional Analysis*, vol. 274, no. 11, pp. 3090–3123, 2018.

[18] Matthias Liero, Alexander Mielke, and Giuseppe Savaré, "Optimal Entropy-Transport problems and a new Hellinger-Kantorovich distance between positive measures," *Inventiones mathematicae*, vol. 211, no. 3, pp. 969–1117, 2018.

[19] Kilian Fatras, Thibault Sejourne, Rémi Flamary, and Nicolas Courty, "Unbalanced minibatch optimal transport; applications to domain adaptation," in *ICML*, 2021.

[20] Aude Genevay, Gabriel Peyré, and Marco Cuturi, "Learning generative models with sinkhorn divergences," in *AISTATS*, 2018.

[21] Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty, "Learning with minibatch Wasserstein: Asymptotic and gradient properties," in *AISTATS*, 2020.

[22] Luis Caffarelli and Robert McCann, "Free boundaries in optimal transport and Monge-Ampère obstacle problems," *Annals of Mathematics*, vol. 171, no. 2, pp. 673–730, 2010.

[23] Ofir Pele and Michael Werman, "Fast and robust Earth Mover's Distances," in *IEEE ICCV*, 2009.

[24] Debarghya Mukherjee, Aritra Guha, Justin Solomon, Yuekai Sun, and Mikhail Yurochkin, "Outlier-Robust Optimal Transport," in *ICML*, 2020.

[25] Ronald R. Coifman and Mauro Maggioni, "Diffusion wavelets," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 53–94, 2006.

[26] William Leeb and Ronald Coifman, "Hölder–Lipschitz Norms and Their Duals on Spaces with Semigroups, with Applications to Earth Mover's Distance," *Journal of Fourier Analysis and Applications*, vol. 22, no. 4, pp. 910–953, 2016.

[27] Longqi Liu, Chuanyu Liu, Andrés Quintero, Liang Wu, Yue Yuan, Mingyue Wang, Mengnan Cheng, Lizhi Leng, Liqin Xu, Guoyi Dong, Rui Li, Yang Liu, Xiaoyu Wei, Jiangshan Xu, Xiaowei Chen, Haorong Lu, Dongsheng Chen, Quanlei Wang, Qing Zhou, Xinxin Lin, Guibo Li, Shiping Liu, Qi Wang, Hongru Wang, J. Lynn Fink, Zhengliang Gao, Xin Liu, Yong Hou, Shida Zhu, Huanming Yang, Yunming Ye, Ge Lin, Fang Chen, Carl Herrmann, Roland Eils, Zhouchun Shang, and Xun Xu, "Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity," *Nature Communications*, vol. 10, no. 1, pp. 470, 2019.

[28] Alan R. Aronson and François-Michel Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.