Diffusion Distance for Signals Supported on Networks

Weiyu Huang, Santiago Segarra, and Alejandro Ribeiro

***Abstract*—We introduce the diffusion distance as a metric to compare signals supported in the nodes of a network. The metric considers the given vectors as initial temperature distributions and diffuses heat through the edges of the graph. The similarity between the given vectors is determined by the similarity of the respective diffusion profiles. The diffusion distance computes the accumulated difference between the diffused signals. We prove that diffusion distance defines a valid metric and is stable to perturbations in the underlying network. We utilize numerical experiments to illustrate its utility in classifying ovarian cancer histologies using gene mutation profiles of different patients. It is also used in a label propagation method in semi-supervised learning to classify handwritten digits.**

I. Introduction

Networks, or graphs, are data structures that encode relationships between elements of a group and which, for this reason, play an important role in many disparate disciplines such as biology [1] and sociology [2] where relationships between, say, genes, species or individuals, are central. Often, networks have intrinsic value and are themselves the object of study. This is the case, e.g., when we are interested in distributed and decentralized algorithms in which agents iterate through actions that use information available either locally or at adjacent nodes to accomplish some sort of global outcome [3]. Equally often, the network defines an underlying notion of proximity, but the object of interest is a signal defined on top of the graph. This is the matter addressed in the field of graph signal processing, where the notions of frequency and linear filtering are extended to signals supported on graphs [4]. Examples of network-supported signals include gene expression patterns defined on top of gene networks [5] and brain activity signals supported on top of brain connectivity networks [6]. Indeed, one of the principal uses of networks of gene interactions is to determine how a change in the expression of a gene, or a group of genes, cascades through the network and alters the expression of other genes. Likewise, a brain connectivity network specifies relationships between areas of the brain, but it is the pattern of activation of these regions that determines the mental state of the subject.

In this paper we consider signals supported on graphs and address the challenge of defining a notion of distance between these signals that incorporates the structure of the underlying network. We want these distances to be such that two signals are deemed close if they are themselves close – in the examples in the previous paragraph we have gene expression or brain activation patterns that are similar –, or if they have similar values in adjacent or nearby nodes – the expressed genes or the active areas of the brain are not similar but they effect similar changes in the gene network or represent activation of closely connected areas of the brain. We define here the diffusion distance and argue that it inherits this functionality through their connection to diffusion processes.

Diffusion processes draw their inspiration from the diffusion of heat through continuous matter [7]. The linear differential equation that models heat diffusion can be extended to encompass dynamics through discrete structures such as networks [8]. In the particular case of networks, every node is interpreted as containing an amount of heat which flows from hot to cold nodes. The flow of heat is through the edges of the graph and such that the rate at which heat diffuses is proportional to both the heat difference between the nodes adjacent to the edge and the edge weight representing the proximity between these nodes. Diffusion processes are often used to exploit their asymptotic configurations in steady state such as in problems of formation control [9] as well as the propagation of opinions in social networks [10].

In this paper we do not exploit the asymptotic, but rather the transient behavior of diffusion processes. We regard the given vectors as initial heat configurations that generate different diffused heat profiles over time. The diffusion metric integrates each of the heat profiles over time and evaluates the norm of the difference between the two integrals. It yields small values when the diffusion profiles are similar. This happens if the given vectors themselves are close or if they have similar values at nodes that are linked by edges with high similarity values.

II. Preliminaries

A. Graphs and networks

We consider networks (or graphs) as triplets $G = (V, E, W)$, where $V = \{1, \ldots, n\}$ is a finite set of n nodes or vertices, $E \subseteq V \times V$ is a set of edges defined as ordered pairs (i, j), and $W : E \to \mathbb{R}_{++}$ is a map from the set of edges to the strictly positive reals, representing weights $w_{ij} > 0$ associated with each edge (i, j). We assume undirectedness where edge $(i, j) \in E$ if and only if $(j, i) \in E$ and symmetry with $w_{ij} = w_{ji}$ for all $(i, j) \in E$. The edge (i, j) represents the existence of a relationship between i and j and we say that i and j are adjacent or neighboring. The weight $w_{ij} = w_{ji}$ represents the strength of the relationship between i and j. Larger edge weights are interpreted as higher similarity between the border nodes. The graphs considered here do not contain self loops, i.e., $(i, i) \notin E$ for any $i \in V$.

We consider the usual definitions of the adjacency, Laplacian, and degree matrices for the weighted graph $G = (V, E, W)$; see e.g. [11, Chapter 1]. The adjacency matrix $A \in \mathbb{R}_+^{n\times n}$ is such that $A_{ij} = w_{ij}$ whenever $(i, j) \in E$ and $A_{ij} = 0$ otherwise. The degree matrix $D \in \mathbb{R}_+^{n\times n}$ is a diagonal matrix with diagonal elements $D_{ii} = \sum_j w_{ij}$ containing the sum of all the weights out of node i. The Laplacian matrix is defined as the difference $L := D - A \in \mathbb{R}^{n\times n}$. Since D is diagonal and the diagonal of A is null, the components of the Laplacian matrix are explicitly given by

$$L_{ij} := \begin{cases} -A_{ij} & \text{if} \quad i \neq j, \\ \sum_{k=1}^{n} A_{ik} & \text{if} \quad i = j. \end{cases} \tag{1}$$

B. Metrics and norms

Our goal in this paper is to define a metric to compare vectors defined on top of a graph. For reference, recall that for a given space X, a metric $d : X \times X \to \mathbb{R}_+$ is a function from pairs of elements in X to the nonnegative reals satisfying the following three properties for every $x, y, z \in X$:

Symmetry: $d(x, y) = d(y, x)$.
Identity: $d(x, y) = 0$ if and only if $x = y$.
Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$.

A closely related definition is that of a norm. In this case we need to have a given vector space Y and consider elements $v \in Y$. A norm $\|\cdot\|$ is a function $\|\cdot\| : Y \to \mathbb{R}_+$ from Y to the nonnegative reals such that, for all vectors $v, w \in Y$ and scalar constant β, it satisfies:

Positiveness: $\|v\| \geq 0$ with equality if and only if $v = \vec{0}$.
Positive homogeneity: $\|\beta\, w\| = |\beta|\, \|w\|$.
Subadditivity: $\|v + w\| \leq \|v\| + \|w\|$.

Work in this paper is supported by NSF CCF-1217963. The authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania, 200 South 33rd Street, Philadelphia, PA 19104. Email: {whuang, ssegarra, aribeiro}@seas.upenn.edu.

Norms are more stringent than metrics because they require the existence of a null element with null norm. However, whenever a norm is defined on a vector space Y it induces a distance in the same space as we formally state next [12, Chapter 1].

Lemma 1 *Given any norm $\|\cdot\|$ on some vector space Y, the function $d : Y \times Y \to \mathbb{R}_+$ defined as $d(r, s) := \|r - s\|$ for all pairs $r, s \in Y$ is a metric.*

C. Diffusion dynamics

Consider an arbitrary graph $G = (V, E, W)$ with Laplacian matrix L and a vector $r = [r_1, \ldots, r_n]^T \in \mathbb{R}^n$ where the component r_i of r corresponds to the node i of G. For a given constant $\alpha > 0$, define the time-varying vector $r(t) \in \mathbb{R}^n$ as the solution of the linear differential equation

$$\frac{\mathrm{d}\, r(t)}{\mathrm{d}\, t} = -\alpha\, L\, r(t), \qquad r(0) = r. \tag{2}$$

The differential equation in (2) represents heat diffusion on the graph G because $-L$ can be shown to be the discrete approximation of the continuous Laplacian operator used to describe the diffusion of heat in physical space [8]. The given vector $r = r(0)$ specifies the initial temperature distribution and $r(t)$ represents the temperature distribution at time t. The constant α is the thermal conductivity and controls the heat diffusion rate. Larger α results in faster changing $r(t)$. The solution of (2) is given by

$$r(t) = e^{-\alpha\, L\, t}\, r, \tag{3}$$

where, for an arbitrary matrix $A \in \mathbb{R}^{n \times n}$, the matrix exponential e^A is defined as

$$e^A := \sum_{k=0}^{\infty} \frac{1}{k!} A^k. \tag{4}$$

The expression in (3) allows us to compute the temperature distribution at any point in time given the initial heat configuration r and the structure of the underlying network through its Laplacian L. Notice that as time grows, $r(t)$ settles to an isothermal equilibrium if the graph is connected.

It is instructive to rewrite (2) componentwise. If we focus on the i-th component of $r(t)$ and use the definition of L in (1) to replace $L_{ik} = -A_{ik}$ and $L_{ii} = \sum_{k=1}^{n} A_{ik}$, it follows that (2) implies

$$\frac{\mathrm{d}\, r_i(t)}{\mathrm{d}\, t} = \sum_{j=1}^{n} \alpha\, A_{ij}\, (r_j(t) - r_i(t)) \,. \tag{5}$$

Further recalling that $A_{ij} = 0$ if i and j are not adjacent and that $A_{ij} = w_{ij}$ otherwise, we see that the sum in (5) entails multiplying each of the differences $r_j(t) - r_i(t)$ between adjacent nodes by the corresponding proximities w_{ij} on top of the constant thermal conductivity α. Thus, (5) is describing the flow of heat through edges of the graph. The flow of heat on an edge grows proportionally with the temperature differential $r_j(t) - r_i(t)$ as well as with the proximity w_{ij}. Nodes with larger proximity tend to equalize their temperatures faster, other things being equal. In particular, two initial vectors $r(0) = r$ and $s(0) = s$ result in similar temperature distributions across time if they are themselves similar – all r_i and s_i components are close –, or if they have similar initial levels at nodes with larger proximity – each component r_i need not be similar to s_i itself but might be similar to the component s_j of a neighboring node for which the edge weight w_{ij} is large. This latter fact suggests that the diffused vectors $r(t)$ and $s(t)$ define a notion of proximity between r and s associated with the underlying graph structure. We exploit this observation to define distances between signals supported on graphs in the following two sections.

III. Diffusion distance

Given an arbitrary graph $G = (V, E, W)$ with Laplacian L, an input vector norm $\|\cdot\|$ and two signals $r, s \in \mathbb{R}^n$ defined in the node space V, the diffusion distance $d^L_{\text{diff}}(r, s)$ between r and s is given by

$$d^L_{\text{diff}}(r, s) := \left\| \int_0^{+\infty} e^{-t}\, e^{-\alpha\, L\, t} (r - s)\, \mathrm{d}t \right\|, \tag{6}$$

with $\alpha > 0$ corresponding to the diffusion constant in (2). As we mentioned in the discussion following (5), the distance $d^L_{\text{diff}}(r, s)$ defines a similarity between r and s that incorporates the underlying network structure. Indeed, we are looking at the difference between the temperatures $r(t)$ and $s(t)$ at time t, which we then multiply by the dampening factor e^{-t}, integrate over all times, and finally take the norm. An interpretation in terms of heat diffusion is that the diffusion distance compares the difference in the total (discounted) energies that pass trough each node. The total energies are similar if initial temperature distributions r and s are similar, or, if r and s have similar values at similar nodes. The dampening factor gives more relative importance to the differences between $r(t)$ and $s(t)$ for early times. This is necessary because after prolonged diffusion times the network settles into an isothermal equilibrium and the structural differences between r and s are lost.

Notice the integral in (6) can be resolved to obtain a closed solution. To do so, observe that the primitive of the matrix exponential $e^{-t}e^{-\alpha L t} = e^{-(I+\alpha L)t}$ is given by $-(I + \alpha L)^{-1} e^{-(I+\alpha L)t}$ to conclude that (6) is equivalent to

$$d^L_{\text{diff}}(r, s) = \left\| (I + \alpha L)^{-1} (r - s) \right\|. \tag{7}$$

Exploiting the same interpretation, we can define the diffusion norm of a vector $v \in \mathbb{R}^n$ for a given graph with Laplacian matrix L and a given input norm $\|\cdot\|$ as

$$\|v\|^L_{\text{diff}} := \left\| \int_0^{+\infty} e^{-t}\, e^{-\alpha\, L\, t} v\, \mathrm{d}t \right\| = \left\| (I + \alpha L)^{-1} v \right\|, \tag{8}$$

The diffusion distance is a proper metric and the diffusion norm is a proper norm. We show first that $\|\cdot\|^L_{\text{diff}}$ is a valid norm as we formally state next.

Proposition 1 *The function $\|\cdot\|^L_{\mathit{diff}}$ in* (8) *is a valid norm on $\mathbb{R}^n$ for every Laplacian L and every input norm $\|\cdot\|$.*

Proof: See [13]. ■

From Proposition 1 and Lemma 1 it follows directly that that the diffusion distance defined in (6) is a valid metric as we state next.

Corollary 1 *The function d^L_{diff} in* (6) *is a valid metric on $\mathbb{R}^n$ for every Laplacian L and every input norm $\|\cdot\|$.*

The distance d^L_{diff} incorporates the network structure to compare two signals r and s supported in a graph with Laplacian L. As in the particular case where the edge set E of the underlying graph G is empty, the Laplacian $L = \mathbf{0}$ is identically null and we obtain from (6) that $d^{\mathbf{0}}_{\text{sps}}(r, s) = \|r - s\|$. This is consistent with the fact that when no edges are present, the network structure adds no information to aid in the comparison of r and s and the diffusion distance reduces to the standard distance induced by the input norm. The same effect is obtained when the thermal conductivity α is set to zero.

In order to illustrate the diffusion distance and its difference with the standard vector distances, consider the undirected graph in Figure 1 where the weight of each undirected edge is equal to 1. Define three different vectors supported in the node space and having exactly one component equal to 1 and the rest equal to 0. The vector r has its positive component for node x_1, colored in red, the vector g has its positive for node x_6, colored in green, and the vector y has its positive component for node x_7, colored in yellow.

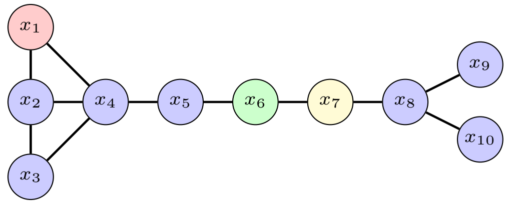
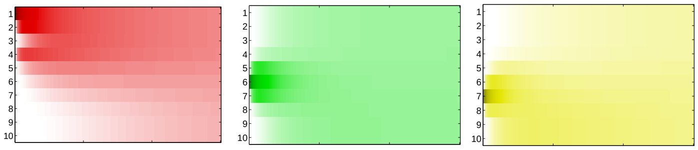


Fig. 1: Example of an underlying graph used to compute the diffusion distance. Three signals r, g and y are compared taking a value of 1 in the red, green, and yellow nodes respectively, and zero everywhere else.



(a) Diffusion of r (b) Diffusion of g (c) Diffusion of y

Fig. 2: Heat maps of the diffused signals for r, g, and y as diffusion evolves for every node in the network in Figure 1. Darker colors represent stronger signals. The heat maps of g and y are more similar, entailing smaller diffusion distance.

For the traditional vector metrics, the distances between each of the vectors r, g and y are the same. In the case when, e.g., the ℓ_2 distance is used as input metric, we have that $\|r-g\|_2 = \|g-y\|_2 = \|y-r\|_2 = \sqrt{2}$. Results are similar in the cases of the ℓ_1 and ℓ_∞ distances. However, by observing the network in Figure 1, it is intuitive that signals g and y should be more alike than they are to r since they affect nodes that are closely related. E.g., if we think of the vectors r, g and y as signaling faulty nodes in a communication network, it is evident that the impact of nodes x_6 and x_7 failing would disrupt the communication between the right and left components of the graph, whereas the failure of x_1 would entail a different effect. This intuition is captured by the diffusion distance. Indeed, if we fix $\alpha = 1$ and we use the ℓ_2 norm as input norm to the diffusion distance, we have that the distance between the vectors that signal faults at x_6 and x_7 are [cf. (7)]

$$d^L_{\text{diff}}(g, y) = \|(I + L)^{-1}(g - y)\|_2 = 0.418, \tag{9}$$

where L is the Laplacian of the graph in Figure 1. However, the diffusion distances from these green and yellow vectors to the red vector that signals a fault at node x_1 are

$$\begin{aligned} d^L_{\text{diff}}(r, g) &= \|(I + L)^{-1}(r - g)\|_2 = 0.664, \\ d^L_{\text{diff}}(r, y) &= \|(I + L)^{-1}(r - y)\|_2 = 0.698. \end{aligned} \tag{10}$$

The distances in (10) are larger than the distance in (9) signaling the relative similarity of the g and y vectors with respect to the r vector. The differences are substantial – almost 60% increase –, thus allowing identification of g and y as somehow separate from r. Further observe that the distance between r and g is slightly smaller than the distance between r and y. This is as it should be, because node x_1 is closer to node x_6 than to node x_7 in the underlying graph.

To further illustrate the intuitive idea behind the diffusion distance, Figure 2 plots the evolution of the diffused signals $r(t)$, $g(t)$ and $y(t)$ for each of the respective initial conditions r, g, and y. At time $t = 0$ each of the signals is concentrated at one specific node. The signals are, as a consequence, equally different to each other. At very long times, the signals are completely diffused and therefore indistinguishable. For intermediate times, the signal distributions across nodes for the green and yellow signals are more similar than between the green and red or yellow and red signals. This difference between the evolution of the diffused signals results in different values for the diffusion distance.

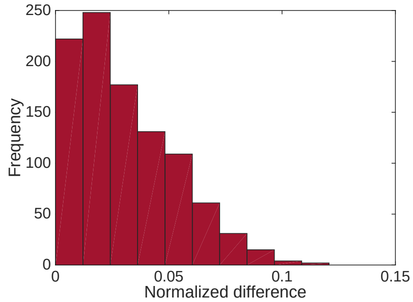


Fig. 3: Histogram of the absolute value of the normalized difference, i.e. $|d^{L'}(g, r) - d^L(g, r)|/\|E\|_2$, for the diffusion distance. For this particular network and perturbations, the difference is considerably lower than the theoretical upper bound of 2.

Remark 1 Computation of the diffusion distance using the closed form expression in (7) requires the inversion of the $n \times n$ identity plus Laplacian matrix followed by multiplication with the difference vector $r - s$. The cost of this computation is of order n^3, but is much smaller when the matrix L is sparse, as is typically the case. Further observe that most computations can be reused when computing multiple distances, because the vectors change, but the matrix inverse $(I + \alpha L)^{-1}$ stays unchanged.

IV. Stability

The diffusion distance depends on the underlying graphs through their Laplacian L. It is therefore important to analyze how a perturbation of the underlying network impacts both distances. We prove in this section that the diffusion distance is well behaved with respect to perturbations of the underlying graph. I.e., we show that if the network perturbation is small, the change in the diffusion distance is also small. We think of a perturbation of a given network as noise added to its edge weights, thus, we quantify the network perturbation as the matrix p-norm of the difference between the Laplacians of the original and perturbed networks. We focus our analysis on the most frequently used norms where $p \in \{1, 2, \infty\}$. The diffusion distance defined in (6) is stable for these input norms as we formally state next.

Proposition 2 *Given any graph with Laplacian L, an input ℓ_p norm $\|\cdot\|_p$ with $p \in \{1, 2, \infty\}$, and bounded signals s and r on the network with $\|s\|_p \leq \gamma$ and $\|r\|_p \leq \gamma$, if we perturb the network such that the resulting Laplacian $L' = L + E$ where the perturbation E is such that $\|E\|_p \leq \epsilon\|L\|_p < 1$, then*

$$\left| d^{L'}_{\textit{diff}}(s, r) - d^L_{\textit{diff}}(s, r) \right| \leq 2\gamma\|L\|_p\epsilon + o(\epsilon). \tag{11}$$

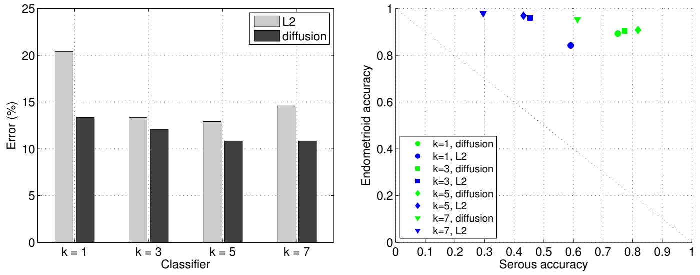
Proof: See [13]. ■

The bound in (11) contains higher order terms that depend on the magnitude of the perturbation. Hence, since the other terms of the bound in (11) tend to zero super linearly, we may divide (11) by $\epsilon\|L\|_p$ and compute the limit as the perturbation vanishes

$$\lim_{\epsilon \to 0} \frac{\left| d^{L'}_{\text{diff}}(s, r) - d^L_{\text{diff}}(s, r) \right|}{\epsilon\|L\|_p} \leq 2\gamma, \tag{12}$$

which implies that vanishing perturbations on the underlying network have vanishing effects on the distance between two signals defined on the network.

When constructing the underlying graph to compare signals in a real-world application, noisy information can be introduced. This means that the similarity weight between two nodes in the underlying graph contains inherent error. Proposition 2 shows that the diffusion distance is impervious to these minor perturbations.



(a) (b)

Fig. 4: Histology classification of ovarian cancer patients based on k nearest neighbors with respect to the ℓ_2 and diffusion distances of their genetic profile. (a) Light bars denote the error when patients are classified using the ℓ_2 distance while the dark bars denote the error when diffusion distance is used for different k-NN classifiers. The diffusion distance reduces the classification error consistently across classifiers. (b) Accuracy of serous subtype vs. endometrioid subtype. Classifiers using diffusion (green) are closer to the top right corner, i.e. perfect classification, than those using the ℓ_2 distance (blue).

In order to illustrate the stability results presented, consider again the underlying network in Figure 1. We perturb this network by multiplying every edge weight by a random number uniformly picked from $[0.95, 1.05]$ and then compute the diffusion distance between vectors r and g with the perturbed graph as underlying network. For these illustrations we pick the input norm to be ℓ_2. In Figure 3 we plot histograms of the absolute value of the difference in the distances when using the original and the perturbed graphs as underlying networks normalized by the norm of the perturbation for 1000 repetitions of the experiment. From (12) we know that this value should be less than 2 for the diffusion distance for vanishing perturbations. Indeed, as can be seen from Figure 3, all perturbations are below the threshold of 2 by a considerable margin. This stability property is essential for the practical utility of the diffusion and superposition distances as seen in the next section.

Remark 2 In Proposition 2 we focus our analysis on the input norms $\|\cdot\|_p$ for $p \in \{1, 2, \infty\}$ because these norms lead to the simple bound in (11). The simplicity of this bound is derived from the fact that $\|(I + L)^{-1}\|_p \leq 1$ for the values of p previously mentioned. For other matrix norms satisfying minor conditions, the equivalence of norms guarantees that bounds analogous to those in (11) must exist, however with potentially more involved constant terms.

V. Applications

We illustrate the advantages of the diffusion distance developed in Section III through numerical experiments in real-world data (Sections V-A and V-B).

A. Ovarian cancer histology classification

We demonstrate that the diffusion distance can provide a better classification of histology subtypes for ovarian cancer patients than the traditional ℓ_2 metric. To do this, we consider 240 patients diagnosed with ovarian cancer corresponding to two different histology subtypes [14]: serous and endometrioid. Our objective is to recover the histology subtypes from patients' genetic profiles.

For each patient i, her genetic profile consists of a binary vector $v_i \in \{0, 1\}^{2458}$ where, for each of the 2458 genes studied, v_i contains a 1 in position k if patient i presents a mutation in gene k and a 0 otherwise. One way of building a metric in the space of 240 patients is by quantifying the distance between patients i and j as the ℓ_2 distance between their genetic profiles,

$$d_{\ell_2}(i, j) = \|v_i - v_j\|_2 \,. \tag{13}$$

In this approach, every gene is considered orthogonal to each other and compared separately across patients. An alternative approach is to take into account the relational information across genes when comparing patients. In order to do so, we apply the diffusion distance on an underlying gene-to-gene network built based on publicly available data [15]. In order to build this network, we first extract the pairwise gene-gene interactions from [15] using the *NCI_Nature* database. After normalization, every edge weight is contained between 0 and 1, which we interpret as a probability of interaction between genes. We assign to each path the probability obtained by multiplying the probabilities in the edges that form the path. For every pair of genes in the network, we compute a similarity value between them corresponding to the maximum probability achievable by a path that links both genes. Finally, we apply normalization and thresholding operations to obtain the gene-to-gene network that we use in our experiments. Observe that the gene-to-gene network contains accepted relations between genes in humans in general and is not patient dependent, hence, it defines a common underlying network for all subjects being compared. Thus, denoting as L the Laplacian of the gene-to-gene network and using the ℓ_2 as input norm we compute the diffusion distances between patients i and j as [cf. (7)]

$$d^L_{\text{diff}}(i, j) = \left\|(I + \alpha L)^{-1}(v_i - v_j)\right\|_2 , \tag{14}$$

where α was set to 15, however, results are robust to this particular choice.

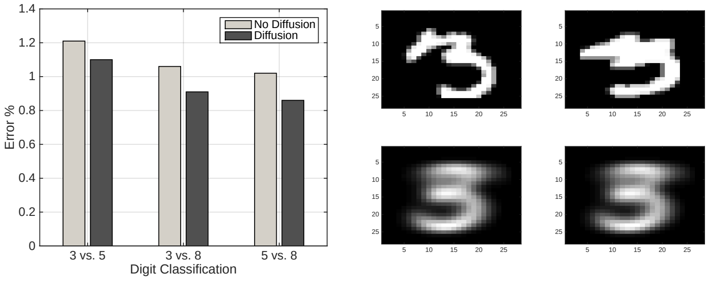
In order to evaluate the classification power of both approaches – ℓ_2 and diffusion distance – we perform 240-fold cross validation for a k nearest neighbors (k-NN) classifier. More precisely, for a particular patient, we look at the k nearest patients as given by the metric being evaluated and assign to this patient the most common cancer histology among the k nearest patients. We then compare the assigned histology with her real cancer histology and evaluate the accuracy of the classifier. Finally, we repeat this process for the 240 women considered and obtain a global classification accuracy for both approaches.

In Figure 4a we show the reduction in histology classification error when using the diffusion distance (14) compared to using the ℓ_2 distance (13) when comparing genetic profiles. The four groups of bars correspond to classifiers built using different numbers of neighbors $k \in \{1, 3, 5, 7\}$. Notice that the reduction in error is consistent across all classifiers analyzed with an average error reduction of over 21%, unveiling the value of incorporating the network information in the classification process.

To further analyze the obtained results, in Figure 4b we present the accuracy obtained for the serous subtype versus the accuracy obtained for the endometrioid subtype for different classifiers based on the diffusion (green) and ℓ_2 (blue) distances. Points on the top right corner of the plot are ideal, obtaining perfect classification for both subtypes. When using diffusion, accuracies shift towards the ideal position since the accuracies for the serous subtypes increase by 20% to 40% whereas the accuracies for endometrioid subtypes decrease by less than 5%. Furthermore, among the 240 patients analyzed, there are 196 of them with endometrioid subtype and only 44 with serous subtype. Hence, a nearest neighbor classifier based on an uninformative distance would tend to have a high classification accuracy for the former but a low one for the latter. This is the case for the ℓ_2 metric. The diffusion distance, in contrast, by exploiting the gene-to-gene interaction can overcome this limitation.

B. Handwritten digit recognition

Diffusion distance can be instrumental in the classification of digits via semi-supervised learning. To illustrate this, consider the well-known MNIST handwritten digit database [16]. Each observation consists of a square gray-scaled image of a handwritten digit with 28 × 28 pixels. Consequently, we can think of each observation as a vector $x \in \mathbb{R}^{784}$ where the value of each component corresponds to the intensity of the associated pixel. A subset of these images – the training set – are labeled, i.e. we know the digit that the image represents. The rest of the images – the testing set – are unlabeled and our objective is to correctly identify



(a) (b)

Fig. 5: Digit recognition based on the traditional and diffused k nearest neighbors approaches. (a) Error rates for three binary classification problems of written digits given by the traditional and diffused k-NN approaches. Error is reduced by diffusion in the three cases. (b) Two instances of handwritten threes (top) which are interpreted as fives by the classical k-NN approach and their corresponding diffused image (bottom). Diffusion averages out irregularities, achieving higher classification accuracy.

the digits they represent. Given n the total number of images – labeled or unlabeled –, we define $X \in \mathbb{R}^{784\times n}$ as $X = [x_1, x_2, \dots, x_n]$ so that each row in X corresponds to the pixels of one digit.

K nearest neighbors is a simple conventional approach used to classify the digits. In order to implement it, we first compute the ℓ_2 pairwise distance between all the vectors x_i. Equivalently, if we denote by e_i the i-th canonical vector – all entries of e_i are zero except the i-th entry which is 1 – the ℓ_2 distance between digits i and j can be written as

$$d_{\ell_2}(i,j) = \left\| X(e_i - e_j) \right\|_2 . \tag{15}$$

To obtain the estimated label of an image in the testing set, we look at the labels of the k closest images among those in the training set as given by (15) and pick the mode of these labels, i.e., the most popular one.

An alternative k-NN approach can be designed using diffusion by defining a graph G_d whose nodes are the handwritten digits. To do this, we draw an edge – with weight 1 – between two digits i and j in G_d if the ℓ_2 pairwise distance (15) is less than a threshold τ. We can interpret digit i as being represented by the signal e_i on G_d, with value 1 at node i and 0 elsewhere. The diffused version of e_i is given by $(I + \alpha L_d)^{-1} e_i$ [cf. (8)] where L_d is the Laplacian of G_d. We can then quantify the distance between two diffused digits i and j as

$$d_{\text{diff}}^{L_d}(i,j) = \left\| X(I + \alpha L_d)^{-1}(e_i - e_j) \right\|_2 . \tag{16}$$

We can then train a k-NN classifier based on the distance between the diffused digits and compare the results with the conventional k-NN based on the ℓ_2 distance without diffusion. Notice that $d_{\text{diff}}^{L_d}(i,j)$ reduces to $d_{\ell_2}(i,j)$ when $L_d = \mathbf{0}$ or when $\alpha = 0$.

In Fig. 5a we present the attribution error comparison between both approaches when performing a binary attribution task between hard-to-distinguish digits: 3 vs. 5, 3 vs. 8, and 5 vs. 8. For each of these cases, we use the entire MNIST training set and testing set with $k \in \{3, 5, 7\}$. It is immediate to see that the diffusion approach outperforms the traditional k-NN in the three tasks. To see why this is the case, in Fig. 5b (top) we present two handwritten images that correspond to threes but are misclassified as fives by the traditional k-NN method. As comparisons, in Fig. 5b (bottom) we present their representations after diffusion in G_d. It is clear that diffusion averages out irregularities found in particular handwritten digits and drives them towards a canonical representation of the number 3.

If we replicate the comparison for a ten class classification problem, i.e. for all digits between 0 and 9, diffusion still improves the accuracy by reducing the error rates from 4.43% to 4.21% (training set of 8600 digits, testing set of 1400 digits and $k = 3$). Moreover, further accuracy improvements can be obtained by combining the traditional and the diffused k-NN methods by choosing the most popular label among the k nearest neighbors in the traditional approach and the $k + 1$ nearest neighbors in the diffused approach. The error rate is further reduced to 3.93%. We pick k neighbors from one approach and $k+1$ from the other to obtain an odd total number of neighbors, reducing the possibility of a multimodal distribution of labels. For the cases where $k \in \{5, 7\}$, similar results are obtained where we see still see the benefit of using diffusion which is further boosted by combining the traditional and the diffused k-NN methods.

Notice that this application of the diffusion distance is fundamentally different from the one presented in Section V-A. In the ovarian cancer case, the nodes in the network represent genes and each signal on the network represents a patient. In contrast, in the current case, both the nodes in the network and the signals represent handwritten digits. This approach can be used in general for label propagation problems in graphs.

VI. Conclusion

We defined the diffusion distance as a metric to compare signals in networks. It relies on the temporal heat map induced by the diffusion of signals across the network and evaluates the accumulated effect across time. We showed the diffusion distance to be stable with respect to perturbations in the underlying network. We demonstrated how diffusion distance can be used to obtain a better classification of signals in a real-world classification of cancer histologies. Finally, we illustrated the use of diffusion as part of a label propagation process to classify handwritten digits.

References

[1] E. Lieberman, C. Hauert, and M. Nowak, "Evolutionary dynamics on graphs," *Nature*, vol. 433, no. 7023, pp. 312–316, 2005.

[2] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E*, vol. 74, p. 036104, 2006.

[3] J. Kleinberg, "Complex networks and decentralized search algorithms," in *Proceedings of the International Congress of Mathematicians (ICM)*, vol. 3, 2006, pp. 1019–1044.

[4] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *Signal Processing Magazine, IEEE*, vol. 30, no. 3, pp. 83–98, May 2013.

[5] R. Mittler, S. Vanderauwera, M. Gollery, and F. V. Breusegem, "Reactive oxygen gene network of plants," *Trends in Plant Science*, vol. 9, no. 10, pp. 490 – 498, 2004.

[6] O. Sporns, *Networks of the Brain*. MIT press, 2011.

[7] E. Eckert and R. Drake, *Analysis of heat and mass transfer*. Hemisphere Publishing; New York, NY, 1987.

[8] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *ICML*, vol. 2, 2002, pp. 315–322.

[9] J. A. F. R. Olfati-Saber and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[10] J. C. Dittmer, "Consensus formation under bounded confidence," *Nonlinear Analysis*, vol. 47, 2001.

[11] F. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.

[12] D. Burago, Y. Burago, and S. Ivanov, *A course in metric geometry*. American Mathematical Society Providence, 2001, vol. 33.

[13] S. Segarra, W. Huang, and A. Ribeiro, "Diffusion and superposition distances for signals supported on networks," *Signal and Information Processing over Networks, IEEE Transactions on*, p. (submitted), 2014.

[14] M. Hofree, J. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nature methods*, 2013.

[15] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. Bader, and C. Sander, "Pathway commons, a web resource for biological pathway data," *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D685–D690, 2011. [Online]. Available: http://nar.oxfordjournals.org/content/39/suppl_1/D685.abstract

[16] Y. Lecun and C. Cortes, "The MNIST database of handwritten digits." [Online]. Available: http://yann.lecun.com/exdb/mnist/