New Resistance Distances with Global Information on Large Graphs

Canh Hao Nguyen[1]   **Hiroshi Mamitsuka**[1,2]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan
[2] Department of Computer Science, Aalto University, Espoo, Finland

Abstract

We consider the problem that on large random geometric graphs, random walk-based distances between nodes do not carry global information such as cluster structure. Instead, as the graphs become larger, the distances contain mainly the obsolete information of local density of the nodes. Many distances or similarity measures between nodes on a graph have been proposed but none are both proved to overcome this problem or computationally feasible even for small graphs. We propose new distance functions between nodes for this problem. The idea is to use electrical flows with different energy functions. Our proposed distances are proved to be metrics in L^p spaces, to keep global information, avoiding the problem, and can be computed efficiently for large graphs. Our experiments with synthetic and real data confirmed the theoretical properties and practical performances of our proposed distances.

1 Introduction

The graph Laplacian is a popular tool to extract information from graphs for various purposes. It is a means to utilize global graph topology for classification with kernels (Smola and Kondor, 2003), graph cut (Shi and Malik, 1997) and clustering (von Luxburg, 2007; Bühler and Hein, 2009). It is particularly useful in semi-supervised learning (Zhou et al., 2003; Belkin et al., 2006; Chapelle et al., 2010) where graphs encode unlabelled data distributions in learning. It is a standard tool in manifold and high dimensional data learning (Belkin and Niyogi, 2003), where only nearest neighbor graphs would give reliable information of

the underlining distributions. Beyond these learning tasks, it also finds applications in other areas such as electrical networks (Klein and Randić, 1993; Bollobás, 1998; I. Gutman, 2004), random walks (Lovasz, 1996) and collaborative filtering (Fouss et al., 2007). The key point is that the graph Laplacians are the information bottleneck to find distributions. It is of great importance to be able to extract the *right* information from the graphs to have the desired distributions.

The common problem is that, given graphs representing unknown distributions of data, the task is to recover the original distributions using graph topologies. In case that graphs encode similarity or neighborhood relationship, a continuously high density region of the original distribution usually contains many well-connected nodes. In order to recognize these nodes belonging to the same region, in the framework of distance or similarity-based learning, these nodes should have small distances to each others (compared to the other nodes of different regions). In graph embedding, these nodes should be close to each other in terms of the distances in the embedding space. Therefore, the desirable distance should contain *global information* of graphs such as cluster structures, bottlenecks and density of the region in between nodes.

For this aim, many similarity and distance functions on graphs use the idea from random walks (Lovasz, 1996; Bollobás, 1998) with the intuition that nodes from the same clusters can be reached by many paths of high probability. However, it is found that, in large graphs, distances based on graph Laplacians differ from the intuition, converging to some noninformative functions (Nadler et al., 2009; von Luxburg et al., 2010, 2014). This is usually attributed to the fact that long random walks tend to forget where they start (the mixing of random walks) (Lovász and Simonovits, 1990). To counter the effect of long random walks (Fouss et al., 2007), many methods give more weights to short walks compared to long ones (Yen et al., 2008; Wu et al., 2012; Chebotarev, 2012). However, these methods are not proved to overcome the problem.

It is proved analytically for resistance distance (also hitting and commute time distances) on random geo-

metric graphs (Penrose, 2003) that as graphs become larger, the distances converge to a function of local density (von Luxburg et al., 2010, 2014), losing global information. This is called the *global information loss problem*. This is a problem for all methods based the graph Laplacians such as **(i)** the graph Laplacian-based embedding and graph kernels (Smola and Kondor, 2003), **(ii)** random walk-based measures (Klein and Randić, 1993; I. Gutman, 2004; Fouss et al., 2007) and **(iii)** spectral clustering (von Luxburg, 2007). Unfortunately, the only known method that provably overcome this problem, p resistance distance (Herbster and Lever, 2009; Alamgir and von Luxburg, 2011), is *computationally infeasible* even for very small graphs.

In this work, we propose two distance functions, named R_p and R_{12} distances, that (1) *provably overcome the global information loss problem* and (2) *are computationally feasible* for large graphs. Our proposed distances would be the first to satisfy both conditions of overcoming the problem and being practically applicable. The idea is to use electrical flows as in resistance distance, as the flows contains all the information from the graph, and can be computed efficiently for all pairs of nodes altogether. To prevent the global information loss problem, different energy functions are proposed, giving higher weights to the global part of the distance functions. Global information of the graph is proved analytically to be kept (in the same manner as p resistance distance). The distances are also proved to be metrics, resulting in embeddings of the graphs into L^p spaces. In our experiments, our proposed distances consistently showed the cluster structures in large graphs while resistance distance failed.

The paper is organized as follows. In Section 2, we analyze the problem of global information loss in random walk-based distances and show how to overcome the problem using variants of resistance distance. We propose the two distance functions for this problem in Section 3 and show their useful properties. We show experimentally the merit of these distances in simulation in Section 4 and real data in Section 5, and then conclude the paper.

2 Problem Setting and Terminology

2.1 Distances on Random Geometric Graphs

Random Geometric Graph. Random geometric graphs are generated from a distribution in some spaces (Bollobás, 1998; Penrose, 2003). They are used to investigate whether statistical estimations on graphs can retain properties of the distributions that are desirable in learning processes such as cluster/class

structures. The objective is that, from the graphs, clusters, classes, and bottlenecks should be detected based solely on the graphs' topologies. Nodes of graphs are sampled according to the distribution (Penrose, 2003). Edges between nodes are generated by connecting either ϵ neighbors or k nearest neighbors. The weight of an edge $w(x_i, x_j)$ determines if it is a similarity graph (e.g. conductance) or dissimilarity graph (e.g. resistance) with different weighting schemes such as: $w(x_i, x_j) = 1$ (unweighted), $w(x_i, x_j) = \|x_i - x_j\|$ or $w(x_i, x_j) = \exp(\|x_i - x_j\|^2/\delta^2)$.

The problem setting is that a random geometric graph $G = (X, E, A)$, where $X = \{x_i\}_{i=1}^n$ is the node set of n nodes drawn randomly from a smooth distribution p (Alamgir and von Luxburg, 2011), and E is the edge set with $\|E\| = m$. Let L^{-1} be the pseudo-inverse of $L = D - A$, which is the graph Laplacian, with D denoting the degree matrix with node degrees on the diagonal, and A is the edge weight matrix.

Distance on Graph. A distance function d on graph G is defined as $d : (X, X) \to R$. For any pair of nodes (x_s, x_t), some examples are:

- $sp(x_s, x_t)$: shortest path distance between x_s and x_t.

- $ht(x_s, x_t)$: hitting time distance: the expected time passage from x_s to arrive at x_t.

- $ct(x_s, x_t) = ht(x_s, x_t) + ht(x_t, x_s)$: commute time distance.

- $res(x_s, x_t) = \frac{1}{2m}ct(x_s, x_t)$: resistance distance.

- $pres(x_s, x_t)$: p resistance distance generalizing resistance distance (Alamgir and von Luxburg, 2011).

Flow and Energy on Graph. Many distances and their properties are derived from flows from one node to another on a graph defined as follows.

- Y: a *flow* is a set of real values assigned to edges of the graph: $Y \in R^{|E|}$ (Bollobás, 1998).

- A unit flow from x_s to x_t is a flow with a total flow 1 coming in at x_s and going out at x_t.

- $I_p(x_s, x_t)$: p *flow* from x_s to x_t satisfying: $I_p(x_s, x_t) \overset{\text{def}}{=} \arg\min_Y \sum_{e \in E} r_e |y_e|^p$ for $p \in R, p \geq 1$ and $Y = \{y_e\}_{e \in E}$ being any unit flow from x_s to x_t (Alamgir and von Luxburg, 2011; Herbster and Lever, 2009).

- $I_2(x_s, x_t)$: (2 flow) the electrical flow from x_s to x_t.

- $E_p : R^{|E|} \to R$: p *energy* function. For any flow Y:

$$E_p(Y) \stackrel{\text{def}}{=} \sum_{e \in E} r_e |y_e|^p. \qquad (1)$$

p energy function is related to the distances as follows: $res(x_s, x_t) = E_2(I_2(x_s, s_t))$ and $pres(x_s, x_t) = E_p(I_p(x_s, x_t))$ as I_p-s are unit flows (Bollobás, 1998; Alamgir and von Luxburg, 2011). From here on, we use energy (of unit flow) to refer to resistance distance.

We define global and local parts of energy function E_p ($: R^{|E|} \to R$) for later proofs as follows. For (x_s, x_t):

- E^{local}: the set of edges incident to either node in the pair: $E^{local}(x_s, x_t) \stackrel{\text{def}}{=} \{e\}_{x_s \in e, e \in E} \cup \{e\}_{x_t \in e, e \in E}$.

- E^{global}: the set of remaining edges: $E^{global}(x_s, x_t) \stackrel{\text{def}}{=} E \setminus E^{local}(x_s, x_t)$.

- E_p^{local}: *local energy function* of E_p. For any $q \in R$:

$$E_p^{local}(I_q(x_s, x_t)) \stackrel{\text{def}}{=} \sum_{e \in E^{local}(x_s, x_t)} r_e |I_q(x_s, x_t)_e|^p.$$

- E_p^{global}: *global energy function* of E_p. For any $q \in R$:

$$E_p^{global}(I_q(x_s, x_t)) \stackrel{\text{def}}{=} \sum_{e \in E^{global}(x_s, x_t)} r_e |I_q(x_s, x_t)_e|^p.$$

According to definitions of local and global set, E_p^{local} and E_p^{global} are local and global parts of energy function E_p and $E_p = E_p^{local} + E_p^{global}$. Please note that this is a slightly different version from the one defined in (von Luxburg et al., 2010; Alamgir and von Luxburg, 2011) without changing its results.

2.2 Global Information Loss problem

Global information loss problem is that distances or similarity measures do not contain the global information of graphs such as clusters or bottlenecks. Global information is necessary for learning purposes such as clustering and classification. Even though this phenomenon is known experimentally for large graphs (Nadler et al., 2009), it can be proved rigorously only for resistance distance on large random geometric graphs. Resistance distance does not contain the information of the area between the nodes (global information), regardless whether between them is a low or high density region (von Luxburg et al., 2010; Alamgir and von Luxburg, 2011):

$$\lim_{n \to \infty} \frac{E_2^{local}}{E_2^{global}} = \infty, \lim_{n \to \infty} \frac{E_2(I_2(x_s, x_t))}{1/d_s + 1/d_t} = 1. \qquad (2)$$

where d_s and d_t are the degrees of node x_s and x_t. As resistance distance only contains information of local density, it has global information loss problem. A demonstration is shown in Figure (1).

3 R_p and R_{12} distances

In this section, we propose two distance functions that give higher weights to the global part of total energy to reflect global information of the graph. We then show their metric, embedding properties, computational analysis, and phase transition by changing parameters.

3.1 Definition

The desideratum is that the two nodes should have a long distance if they are from different clusters. Therefore, the distances should reflect not only geodesic distance (like shortest path distance), but also the density of the space in between the two nodes. Following the same principle as p resistance distance (Alamgir and von Luxburg, 2011), we wish to design distances that can be decomposed into global and local parts. To reflect global information such as clusters, the distances should have the global part at least of the same magnitude as the local part.

The global information loss problem on large graphs occurs due to *two issues*. The first issue is that large graphs divide a unit flow into many paths, resulting in small flows on edges in the global part ($y_e > y_{e'}$ for $e \in E^{local}, e' \in E^{global}$). The second issue is that electrical energy function $E_2(I_2) = \sum r_e y_e^2$, quadratic on flows on edges, further reduces the weights of energy on the global parts ($r_e y_e^2 >> r_e' y_{e'}^2$).

Our solution to the global information loss problem of resistance distance is to address the *second issue*, electrical energy function. We propose to use E_p for $p < 2$ to define energies (distances)[1] based on electrical flow $I_2(x_s, x_t) \in R^{|E|}$. E_p with $p < 2$ can give a higher weight to the global part of the total energy. That is, for smaller p, the difference between $r_e y_e^2$ and $r_e' y_{e'}^2$ will not be amplified as much as $p = 2$. We propose two distances: R_p *distance* and R_{12} *distance* as follows.

Definition 3.1. *(R_p distance).*

$$R_p(x_s, x_t) \stackrel{def}{=} (E_p(I_2(x_s, x_t)))^{\frac{1}{p}}$$
$$= (\sum_{e \in E} r_e |I_2(x_s, x_t)_e|^p)^{\frac{1}{p}}. \qquad (3)$$

---

[1]For unit flows, resistance distance coincides with energy.

Definition 3.2. *(R_{12} distance) For a given α:*

$$R_{12}(x_s, x_t) \stackrel{def}{=} \alpha R_2(x_s, x_t) + (1-\alpha)R_1(x_s, x_t) \quad (4)$$
$$= \alpha(\sum_{e \in E} r_e |I_2(x_s, x_t)_e|^2)^{\frac{1}{2}} + (1-\alpha)\sum_{e \in E} r_e |I_2(x_s, x_t)_e|.$$

One unique feature of our distances is that, we can use the information of I_2 in the same way as sparsity-inducing norms in Optimization. The distances have parameters p and α to interpolate, containing $R_2 = \sqrt{E_2(I_2)}$ and $R_1 = E_1(I_2)$ in their spectra. For $p \to 1$ and $\alpha \to 0$, they both converge to R_1: the total flows on all edges added together. They have interesting properties as follows.

- Consider the case that flows between the two nodes can be separate paths as in Figure 2. $E_1(I_2)$ is the harmonic mean of the lengths of all paths. It is equal to the shortest path added with the amount of flow outside of the shortest paths (the shaded region).

- R_p and R_{12} distances are different from p resistance distance, randomized shortest path distance and others in the sense that they do not converge to shortest path distance. In all spectra of their parameters, they contain the information of connectivity in between the two nodes present in I_2 (taking into account all paths, instead of only the shortest path).

3.2 Dealing with Global Information Loss

Intuitively, $E_1(I_2)$ energy function overcomes the problem of $E_2(I_2)$ by giving more weights to global part. Any s, t min cut would contain the edges with the total flow weight of 1, having total energy of the same order of magnitude as that of the local part of $E_1(I_2)$ (containing only one s, t min cut). Hence, the global part, consisting of many disjoint s, t min cuts, dominates the local one. We prove that, in $E_1(I_2)$, the global energy dominates the local energy for ϵ-neighborhood and knn graphs as follows.

Theorem 3.3. *For connected ϵ-neighborhood random geometric graphs constructed from a valid region X in R^d (von Luxburg et al., 2014), the global part of $E_1(I_2)$ ($E_1^{global}(I_2)$) dominates the local part ($E_1^{local}(I_2)$) almost surely (for any pair (x_s, x_t)) as $n \to \infty$. Concretely, the following statements hold:*

1. *For unweighted graph $w_{ij} = 1$: $\lim_{n \to \infty} \frac{E_1^{global}}{E_1^{local}} \to \infty$ almost surely as $n \to \infty$ and $\epsilon \to 0$.*

2. *For Euclidean weighted graph with $w_{ij} = d(x_i, x_j)$: $\frac{E_1^{global}}{E_1^{local}} \to \infty$ almost surely as $n \to \infty$ and $\epsilon \to 0$.*

3. *For Gaussian weighted graph with $w_{ij} = \exp(\frac{d(x_i, x_j)^2}{\delta^2})$: $\frac{E_1^{global}}{E_1^{local}} \to \infty$ almost surely as $n \to \infty$, $\epsilon \to 0$ and $O(\delta) > O(\frac{\epsilon}{\sqrt{-\ln(\epsilon)}})$.*

Theorem 3.4. *For connected k-nearest neighbor (random geometric) graphs constructed from a valid region X in R^d (von Luxburg et al., 2014), the global part of $E_1(I_2)$ dominates the local part almost surely as $n \to \infty$. Concretely, there exist constants c_1, c_2 that the following statements hold:*

1. *For unweighted graph $w_{ij} = 1$: $\lim_{n \to \infty} \frac{E_1^{global}}{E_1^{local}} \to \infty$ almost surely as $n \to \infty$, $k > \log(n)$ and $\frac{k}{n} \to 0$ with a probability of at least $1 - c_1 n \exp(-c_2 \sqrt{nk})$ (converging to 1).*

2. *For Euclidean weighted graph with $w_{ij} = d(x_i, x_j)$: $\frac{E_1^{global}}{E_1^{local}} \to \infty$ almost surely as $n \to \infty$, $k > \log(n)$ and $\frac{k}{n} \to 0$ with a probability of at least $1 - c_1 n \exp(-c_2 \sqrt{nk})$ (converging to 1).*

3. *For Gaussian weighted graph with $w_{ij} = \exp(\frac{d(x_i, x_j)^2}{\delta^2})$: $\frac{E_1^{global}}{E_1^{local}} \to \infty$ almost surely as $n \to \infty$, $k > \log(n)$, $\frac{k}{n} \to 0$ and $O(\delta) = (\frac{k}{n})^{\frac{1}{d}}$ with a probability of at least $1 - c_1 n \exp(-c_2 k \cdot \log(\frac{n}{k})^{\frac{d}{2}})$ (converging to 1).*

Please find the proofs in the supplementary material. The valid region definition is the same as in (von Luxburg et al., 2014), but in fact, we only require that distribution p is bounded and the graphs are connected. Note that the conditions of parameters in these theorems cover all the conditions for hitting times and commute distances to lose global information (von Luxburg et al., 2014). Hence, our proposed distance can retain global information in the cases that resistance distance cannot.

3.3 L^p Space Embedding of Graphs

We show that R_p and R_{12} distances are metrics in corresponding L^p spaces. Therefore, the distances naturally lead to embeddings of the graphs onto L^p spaces that preserve global information. The idea is to utilize the formulation of I_2 from L^{-1}. Let $e_i \in R^n$ denote a vector of all zeroes except for the i-th position having value 1.

Lemma 3.5. *Let $V^{(i)} = L_{.i}^{-1}$ denote the i-th column of L^{-1}, respectively. Then,*

$$V^{(s)} - V^{(t)} = L^{-1}(e_s - e_t) \quad (5)$$

is a possible potential assignment to the nodes of the network that makes the unit flow from x_s to x_t.
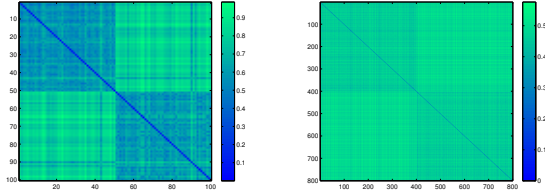
Figure 1: Heatmaps of pairwise resistance distances for a two-cluster distribution (details in Section 4.1) with different numbers of data points sampled: 100 points (left) and 800 points (right). Block structures of the distance matrices, indicating cluster structures, were lost for large graphs.
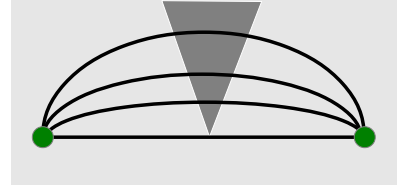

Figure 2: Composition of $E_1(I_2)$ in a simple network: shortest path distance (total flow in non-shaded region) plus the amount of flow on longer paths (shaded region).

Theorem 3.6. *(L^p space embedding) The following embedding f of the nodes of graph G into an L^p space:*

$$f : X \to R^{|E|}$$

$$x_s \to f(x_s) = \{\cdots, \frac{V_i^{(s)} - V_j^{(s)}}{r_{ij}^{(p-1)/p}}, \cdots\}_{(i,j)\in E}^T \quad (6)$$

makes the p-norm of the space coincide with R_p:
$\|f(x_s) - f(x_t)\|_p = R_p(x_s, x_t).$

Proof in the supplementary file for brevity.

Corollary 3.7. *Both R_p and R_{12} are metrics.*

R_{12} is also a metric because it is a convex combination of two metrics.

3.4 Computing R_p and R_{12} distances

We show that these distances can be computed efficiently for all pairs of nodes in large graphs, as opposed to p resistance distance even in small graphs.

Theorem 3.8. *Given the inverse Laplacian matrix L^{-1}, the R_p and R_{12} distances between any pair of nodes can be computed in $O(m)$ time.*

Proof. The formulas (5) and (6) allow us to compute R_p and R_{12} distances in $O(m)$ time complexity from L^{-1} as computing the p-norm in an m dimensional space. Therefore, the total cost of computing all pairwise distances of a graph is the time to inverse the matrix L plus $O(mn^2)$. $\square$

This is different from p resistance distance, which is obtained by solving an $O(n^2)$ number of optimization problems, each with m variables and n linear equality constraints.

3.5 The phase transition

R_p and R_{12} have trade-off parameters p and α respectively. When $p = 1$ and $\alpha = 0$, the global energy

dominates the local one as both distances converge to $R_1(I_2)$. When $p = 2$ and $\alpha = 1$, the local one dominates global ones as both distances converge $R_2(I_2)$. Due to the the continuity of the distances as functions of p and α, there should be phase transition points p^* and α^* that separate the dominant sides: local or global. The exact transition points are hard to compute (Alamgir and von Luxburg, 2011), but can be estimated roughly. For example, let $\beta^* = 1 - \alpha^*$, then, with the usual definition of local and global parts as in (Alamgir and von Luxburg, 2011), the transition point can be characterized as follows.

$$O(\beta^*) = O(\frac{E_2(I_2)^{local}}{E_1(I_2)^{global}}) \quad (7)$$

The phase transition points are necessary in the cases that we wish to have local information having the same or higher weight compared to global one.

3.6 Comparison with p Resistance Distance

Our proposed distances and p resistance distance are different in the flows used. p resistance distance uses I_p to address the *first issue* of $E_2(I_2)$. However, for all pairs of nodes (x_s, x_t), $I_p(x_s, x_t)$ is computed independently by solving an optimization problem, making this method computationally infeasible even for a small. On the other hand, our proposed distances use I_2, which can be *computed for all pairs of nodes altogether*, are computationally efficient for large graphs.

Our proposed distances and p resistance distance are similar in energy function E_p to address the second issue of $E_2(I_2)$. For small p (< 2), E_p gives more weight to the global part of corresponding flows (I_2 and I_p), making E_p^{global} larger. This is key to have global information not to be dominated by local information. Fortunately, addressing the second issue is enough to overcome the global information loss problem.

4 Simulation

We showed simulated experiments in which graphs were generated with known cluster structures to show that our distances can overcome the global information loss problem. The distances being compared were: (i) R_2 (square roots of resistance distances), (ii) R_1, (iii) R_p with $p = \frac{d+1}{d}$, (iv) R_{12} with $\alpha = 1 - \frac{1}{k \cdot n^{1/d}}$, and (v) the shortest path distance (sp). These values of parameters p and α were set to be around the transition points (these are recommended values). Note that, p resistance distance, the only method that provably overcome the problem, cannot be experimented due to its high computational complexity even for very small graphs.

The distances are compared by the quality of their clustering solutions using *clustering accuracy*. Clustering accuracy measures how clustering results agree with the known labels (each label corresponds to a component in the distribution, expecting as a cluster in graph). To compare metric distances for clustering purpose, in all our experiments, we used k-medoids algorithm on these (precomputed) distances and reported the accuracy of the best assignment of clusters to the (given) classes of data. Distances were also visualized in heat maps for intuition (supplementary file).

4.1 Two-cluster case

We designed simulations with clear cluster structures to show that our proposed distances take into account global graph information. The experimental setup was as follows. Mixtures of two anisotropic Gaussian components $f_i(x) = c \cdot \exp(-\frac{1}{2\delta^2}|x - m_i|^2), i = 1, 2$, with $x, m_i \in R^d$ were generated, having standard deviation in any direction of $\delta = 1$. $n/2$ points were sampled independently from each component, then all the n points were used to construct $k \approx \log(n)$ nearest neighbor graphs (random geometric graphs). Data points from the same component are expected form a cluster in graphs, stored in each half of the data set. Distances taking global information of graphs into account should have short distances for the pairs of points in the same half. A demo of the global information loss problem for resistance distance is shown in Figure 1.

4.2 Data size effect

In this simulation, we set $d = 7$, $|m_1 - m_2| = 4$ (being 4 apart from each other) for different sizes of graphs: $n = 50$ to 800, with the corresponding numbers of nearest neighbors of $k = 5$ to 9 (namely, $k \cong \log(n)$).[2]

Clustering accuracy of the distances were shown in Figure 3 (a). We could observe that clustering accuracy of R_2 decreased as n increased. However, clus-
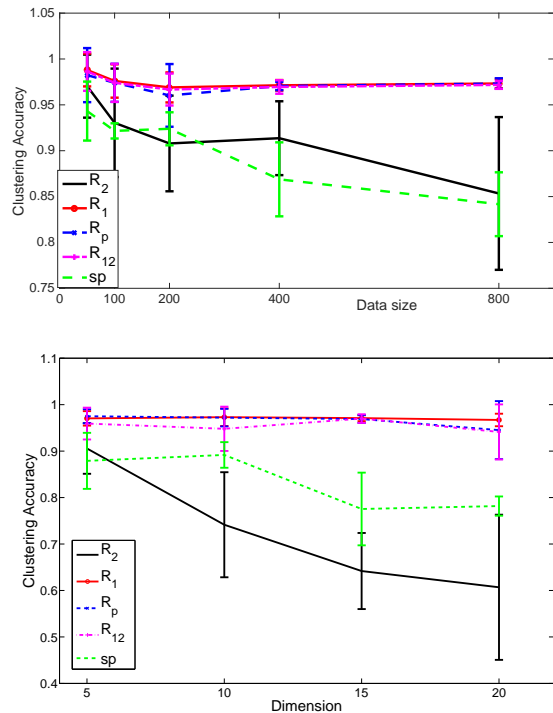

Figure 3: Clustering accuracies for the distances at (a) different data sizes (upper, for dimension 7) and (b) different dimensions (lower, for data size of 200) of the spaces (x axis).

tering accuracies of our proposed distances remained stable and significantly higher than that of R_2. The results of our proposed distances were also much more stable (low variances) compared to that of R_2. This also confirmed the merit of our proposed distances.

4.3 Dimensional effect

We also found experimentally that as the dimension of the space (d) increased (with a fixed data size), resistance distance also lost global information in the same way as data size increased. We showed simulations over different dimensions $d = 5, 10, 15$ and 20 for $n = 200^2$. We showed clustering accuracy for different dimensions in Figure 3 (b). We could observed that, as the dimension became larger, our proposed distances were still able maintained high clustering accuracies while R_2 could not. Our proposed distances were also more stable, having much lower variances. This also confirmed the merit of our proposed distances. However, in too high dimension spaces, nearest neighbor distances are not reliable, making neighborhood graphs and all distances based on these graphs,

---

[2]The heat maps of the distances are shown in the supplementary file for intuition.

Table 1: Clustering accuracies of the distances for different data sets.

	Clustering accuracy (in percentage)			
	R_2	R_1	R_p	R_{12}
idian	69.4±2.9	65.8±2.1	64.7±2.4	65.5±1.7
pop	86.5±8.9	59.8±6.4	62.8±6.4	61.4±9.2
spect	72.2±6.7	57.4±5.0	57.5±3.5	60.3±9.8
ecoli	77.4±10.7	67.9±8.4	74.1±6.4	73.8±9.8
iris	**60.1±9.1**	**87.9±7.5**	**71.9±17.3**	**78.6±7.1**
column	**69.7±0.0**	**73.4±1.7**	**72.6±2.0**	**73.2±1.8**
breast	93.7±10.7	89.6±12.0	79.6±12.2	88.1±1.3
ionosphere	**60.3±2.6**	**65.2±0.0**	**65.2±0.0**	**63.6±4.5**
user	**44.2±3.0**	**49.4±2.4**	**49.1±4.4**	**49.0±4.3**

including our proposed distances, meaningless.

5 Experiments

We showed experiments on real supervised data with known classes[2], to see whether our proposed distances can discover *cluster structures corresponding to classes* more effectively than resistance distance. We selected data sets for classification from UCI Machine Learning repository[3] with numerical attributes and the sizes that the global information loss problem might occur (> 100 nodes). k nearest neighbor graphs were generated with $k > \log(n)$. We grouped nodes of graphs according to their class labels to see if they formed clusters. We showed clustering accuracy for each data set in Table 1. There were *two scenarios*. The first scenario was that the classes formed clear cluster structures, even when resistance distance failed to capture the cluster structures, our distances were expected to show them. The second scenario was that the classes did not form cluster structures, therefore, the graphs did not have clusters corresponding to the classes. In this scenario, it was reasonable to expect that all the distances failed to show cluster structures.

The clustering accuracies indicated that our proposed distances performed well in practice. It could be observed from Table 1 that clustering accuracies of R_2 were never significantly higher than those of our proposed distances. Especially, clustering accuracies of R_2 were significantly lower in *iris, column, ionosphere* and *user* data sets. It is possible that our proposed distances had higher clustering performances than resistance distance (R_2) because they overcame the global information loss problem, as in the first scenario. It was also noteworthy that our proposed distances did not always guarantee performance improvements. This could be the second scenario that the data might not suffer from global information loss problem

that we were targeting. Given that this was supervised data and classes might not coincide with clusters, this result was very encouraging that our distances could show cluster structures.

We show the heat maps of distances on real data to show how clear cluster structures can be seen by using our proposed distances compared to resistance distance in Figure 4.

6 Conclusion

We proposed two distance functions, R_p and R_{12}, that overcome the problem of global information loss in large graphs. Our proposed distances are proved to have global information dominated local one if the parameters are properly set in the same manner as p resistance distance. The distances are metrics, naturally leading to embeddings of graphs into L^p spaces. They can be computed efficiently for large graphs. In synthetic and real data experiments, they were shown to be effective in capturing cluster structures while resistance distance failed even for simple cases. Future work includes elucidating more theoretical properties of edge space embedding, scaling up the distances for massive data sets, transforming from distances to similarity measures such as kernels for classification and other tasks as well as applications.
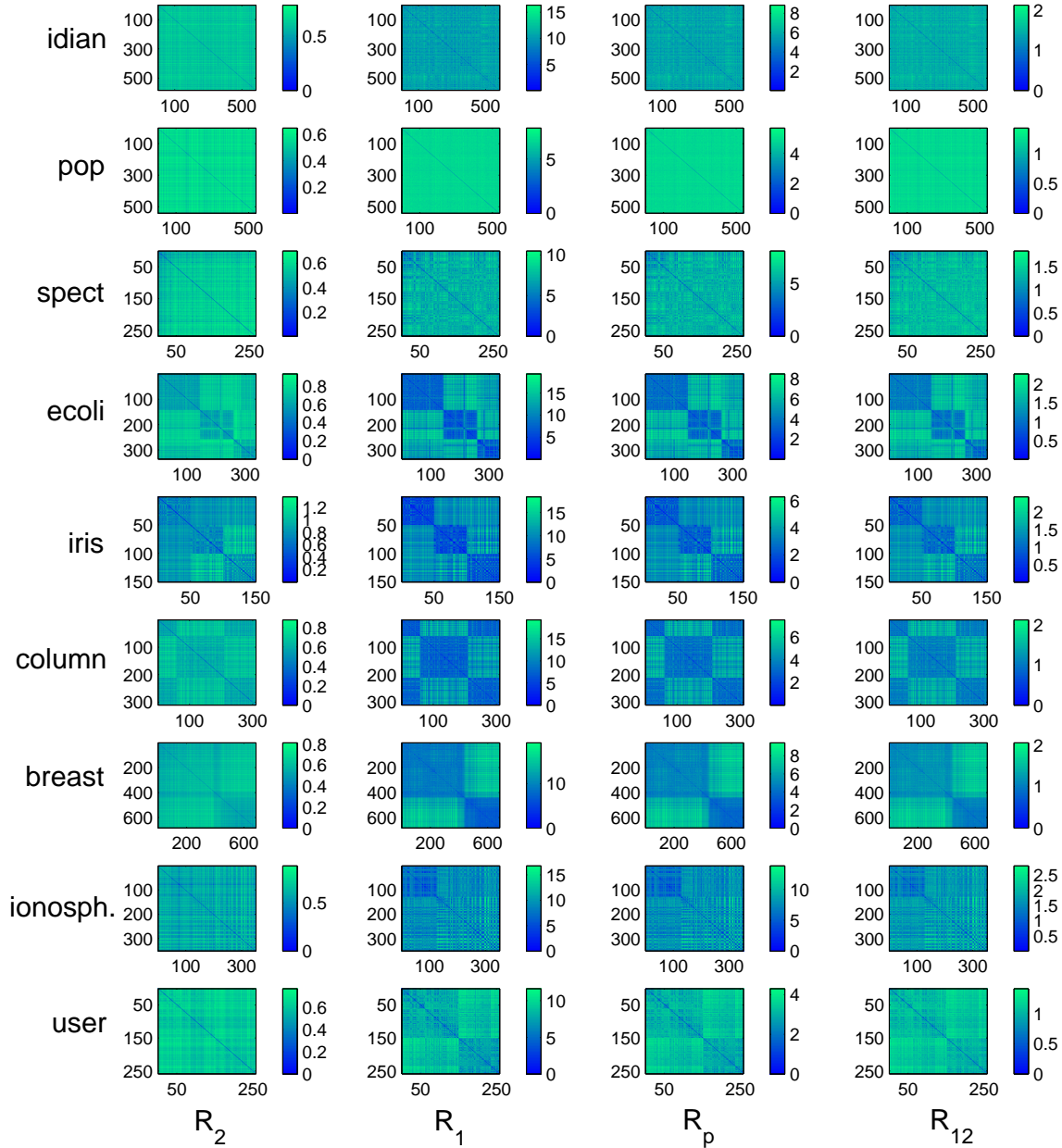
[3]http://archive.ics.uci.edu/ml/

Figure 4: Results on real datasets from UCI Machine Learning repository. The datasets were for classification. We grouped the nodes according to their classes to see whether the distances on the generated graphs could show cluster structures by having diagonal blocks with shorter pairwise distances. The first three datasets showed that no distances could discover cluster structures, probably because the classes did not form clusters. For the rest of the datasets, our distances could always show clear cluster structures. On the other hand, resistance distance either not clearly showed the cluster structures, or completely failed to show them.

References

Morteza Alamgir and Ulrike von Luxburg. Phase transition in the family of p-resistances. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 379–387, 2011.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003. ISSN 0899-7667.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Bélla Bollobás. *Modern Graph Theory*. Springer-Verlag, 1998.

Thomas Bühler and Matthias Hein. Spectral clustering based on the graph p-laplacian. In *ICML*, pages 81–88, 2009.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010. ISBN 0262514125, 9780262514125.

Pavel Chebotarev. The walk distances in graphs. *Discrete Applied Mathematics*, 160(10-11):1484–1500, 2012.

Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowleldge and Data Engineering*, 19(3): 355–369, March 2007.

Mark Herbster and Guy Lever. Predicting the labelling of a graph via minimum p-seminorm interpolation. In *COLT*, 2009.

W. Xiao I. Gutman. Generalized inverse of the laplacian matrix and some applications. *Bulletin. Classe des Sciences MathŐmatiques et Naturelles. Sciences MathŐmatiques*, 129(29):15–23, 2004.

D. J. Klein and M. Randić. Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95, December 1993.

Laszlo Lovasz. Random walks on graphs: A survey. In *Combinatorics, Paul Erdos is Eighty*, volume 2, 1996.

László Lovász and Miklós Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *FOCS*, pages 346–354, 1990.

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *NIPS*, pages 1330–1338, 2009.

Mathew Penrose. *Random Geometric Graphs*. Publisher: Oxford University Press, July 3 2003. ISBN 0198506260.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 1997.

Alex J. Smola and Risi Kondor. Kernels and regularization on graphs. In *COLT*, pages 144–158, 2003.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Ulrike von Luxburg, Agnes Radl, and Matthias Hein. Getting lost in space: Large sample analysis of the resistance distance. In *Advances in Neural Information Processing Systems 23*, pages 2622–2630, 2010.

Ulrike von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15(1):1751–1798, 2014.

Xiao-Ming Wu, Zhenguo Li, Anthony Man-Cho So, John Wright, and Shih-Fu Chang. Learning with partially absorbing random walks. In *NIPS*, pages 3086–3094, 2012.

Luh Yen, Marco Saerens, Amin Mantrach, and Masashi Shimbo. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *KDD*, pages 785–793, 2008.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *NIPS*, 2003.