Kernels on Graphs as Proximity Measures

Konstantin Avrachenkov, Pavel Chebotarev, Dmytro Rubanov

Kernels on Graphs as Proximity Measures

Konstantin Avrachenkov[1], Pavel Chebotarev[2], Dmytro Rubanov[1]

[1] Inria Sophia Antipolis, France
{k.avrachenkov,dmytro.rubanov}@inria.fr
[2] RAS Institute of Control Sciences, Russia
pavel4e@gmail.com

Abstract. Kernels and, broadly speaking, similarity measures on graphs are extensively used in graph-based unsupervised and semi-supervised learning algorithms as well as in the link prediction problem. We analytically study proximity and distance properties of various kernels and similarity measures on graphs. This can potentially be useful for recommending the adoption of one or another similarity measure in a machine learning method. Also, we numerically compare various similarity measures in the context of spectral clustering and observe that normalized heat-type similarity measures with log modification generally perform the best.

1 Introduction

Many graph-based semi-supervised learning methods, see e.g., [1–3, 7, 23, 24, 43] and references therein, can be viewed as the methods comparing some *distances* or *similarity measures* from unlabelled nodes to the labelled ones. An unlabelled node is attributed to a class whose labelled nodes are closer with respect to distances or similarity measures. Also, unsupervised machine learning methods such as K-means and its numerous variations are based on grouping points in a metric space, see e.g. [19, 32, 44]. While the plain K-means method discovers only linear boundaries between clusters in a metric space, kernel K-means methods have better sensitivity and can discover clusters of more general shapes. In addition, some kernel K-means methods are equivalent to spectral clustering [19]. A choice of a kernel may have significant impact on the clustering quality. Moreover, the distance property of the kernels can be exploited for quick grouping of points in the K-means methods [18, 19]. Similarity measures are also used in the link prediction problem [5, 33].

Most but not all similarity measures are defined with the help of kernels on graphs, i.e., positive semidefinite matrices with indices corresponding to the nodes. Note that according to Schoenberg's theorem [34, 35] one can always transform a positive semidefinite matrix to a set of points in an Euclidian space. In contrast, the proximity property [13] is much more subtle and not all kernels on graphs appear to be proximity measures.

In this paper, we analyse distance and proximity properties of the similarity measures and kernels on graphs. All similarity measures and kernels that we

study are defined in terms of one of the following three basic matrices: weighted adjacency matrix, combinatorial Laplacian and (stochastic) Markov matrix. We compare similarity measures and kernels on graphs both theoretically and by numerical experiments in the context of spectral clustering on the stochastic block model. We hope that our analysis will be useful for recommending the adoption of one or another similarity measure in a machine learning method. It was interesting to observe that in the context of the spectral clustering, the normalized heat-type kernels with logarithmic transformation perform the best on the stochastic block model.

2 Definitions and preliminaries

The *weighted adjacency matrix* $W = (w_{ij})$ of a weighted undirected graph G with vertex set $V(G) = \{1, \ldots, n\}$ is the matrix with elements

$$w_{ij} = \begin{cases} \text{weight of edge } (i,j), \text{ if } i \sim j, \\ 0, \qquad\qquad\qquad\quad \text{otherwise.} \end{cases}$$

In what follows, G is connected.

The ordinary (or combinatorial) *Laplacian matrix* L of G is defined as follows: $L = D - W$, where $D = \mathrm{Diag}(W \cdot \mathbf{1})$ is the degree matrix of G, $\mathrm{Diag}(\boldsymbol{x})$ is the diagonal matrix with vector $\boldsymbol{x}$ on the main diagonal, and $\mathbf{1} = (1, \ldots, 1)^T$. In most cases, the dimension of $\mathbf{1}$ is clear from the context.

Informally, given a weighted graph G, a *similarity measure* on the set of its vertices $V(G)$ is a function $\kappa \colon V(G) \times V(G) \to \mathbb{R}$ that characterizes similarity (or affinity, or closeness) between the vertices of G in a meaningful manner and thus is intuitively and practically adequate for empirical applications [2, 18, 24, 33].

A kernel on graph is graph similarity measure that has an inner product representation. Inner product matrices (also called Gram matrices) with real entries are symmetric positive semidefinite matrices. On the other hand, any semidefinite matrix has a representation as a Gram matrix with respect to the Euclidean inner product [25].

We note that following [31, 39] we prefer to write *kernel on graph* rather than *graph kernel*, as the notion of "graph kernel" refers to a kernel between graphs [41].

A *proximity measure* (or simply *proximity*) [13] on a finite set A is a function $\kappa \colon A \times A \to \mathbb{R}$ that satisfies the *triangle inequality for proximities*, viz.:
for any $x, y, z \in A$, $\kappa(x,y) + \kappa(x,z) - \kappa(y,z) \leq \kappa(x,x)$, and if $z = y$ and $y \neq x$, then the inequality is strict.

A proximity κ is a *Σ-proximity* ($\Sigma \in \mathbb{R}$) if it satisfies the *normalization condition*: $\sum_{y \in A} \kappa(x,y) = \Sigma$ for any $x \in A$.

By setting $z = x$ in the triangle inequality for proximities and using the arbitrariness of x and y one verifies that any proximity satisfies *symmetry*: $\kappa(x,y) = \kappa(y,x)$ for any $x, y \in A$.

Furthermore, any Σ-proximity has the *egocentrism* property: $\kappa(x,x) > \kappa(x,y)$ for any distinct $x, y \in A$ [13]. If $\kappa(x,y)$ is represented by a matrix $K = (K_{xy}) = (\kappa(x,y))$, then egocentrism of $\kappa(x,y)$ amounts to the *entrywise diagonal dominance* of K.

If $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are two points in the Euclidean space $\mathbb{R}^n$, then $||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2$ is the squared distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Schoenberg's theorem establishes a connection between positive semidefinite matrices (kernels) and matrices of Euclidean distances.

Theorem 1 ([34, 35]) *Let K be an $n \times n$ symmetric matrix. Define the matrix*

$$\mathcal{D} = (d_{ij}) = \frac{1}{2}\big(\operatorname{diag}(K) \cdot \mathbf{1}^T + \mathbf{1} \cdot \operatorname{diag}(K)^T\big) - K, \tag{1}$$

where $\operatorname{diag}(K)$ *is the vector consisting of the diagonal entries of K. Then there exists a set of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^n$ such that $d_{ij} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2$ $(i, j = 1, \ldots, n)$ if and only if K is positive semidefinite.*

In the case described in Theorem 1, K is the Gram matrix of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Given K, these vectors can be obtained as the columns of the unique positive semidefinite real matrix B such that $B^2 = B^T B = K$. B has the expression $B = U\Lambda^{1/2}U^*$, where $\Lambda = \operatorname{Diag}(\lambda_1, \ldots, \lambda_n)$, $\Lambda^{1/2} = \operatorname{Diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$, and $A = U\Lambda U^*$ is the unitary decomposition of A [25, Corollary 7.2.11].

Connections between proximities and distances are established in [13].

Theorem 2 *For any proximity κ on a finite set A, the function*

$$d(x,y) = \frac{1}{2}(\kappa(x,x) + \kappa(y,y)) - \kappa(x,y), \quad x, y \in A \tag{2}$$

is a distance function $A \times A \to \mathbb{R}$.

This theorem follows from the proof of Proposition 3 in [13].

Corollary 1 *Let $\mathcal{D} = (d_{xy})$ be obtained by* (1) *from a square matrix K. If $\mathcal{D}$ has negative entries or $\sqrt{d_{xy}} + \sqrt{d_{yz}} < \sqrt{d_{xz}}$ for some $x, y, z \in \{1, \ldots, n\}$, then the function $\kappa(x,y) = K_{xy}$, $x, y \in \{1, \ldots, n\}$ is not a proximity.*

Proof. If $\sqrt{d_{xy}} + \sqrt{d_{yz}} < \sqrt{d_{xz}}$, then $d_{xy} + d_{yz} + 2\sqrt{d_{xy}d_{yz}} < d_{xz}$, i.e., the function $d(x,y) = d_{xy}$ violates the ordinary triangle inequality. Thus, it is not a distance, as well as in the case where $\mathcal{D}$ has negative entries. Hence, by Theorem 2, κ is not a proximity. □

The following theorem describes a one-to-one correspondence between distances and Σ-proximities with a fixed Σ on the same finite set.

Theorem 3 ([13]) *Let $\boldsymbol{S}$ and $\boldsymbol{D}$ be the set of Σ-proximities on A ($|A| = n$; $\Sigma \in \mathbb{R}$ is fixed) and the set of distances on A, respectively. Consider the mapping $\psi(\kappa)$ defined by* (2) *and the mapping $\varphi(d)$ defined by*

$$\kappa(x,y) = d(x,\cdot) + d(y,\cdot) - d(x,y) - d(\cdot,\cdot) + \frac{\Sigma}{n}, \tag{3}$$

where $d(x,\cdot) = \frac{1}{n}\sum_{y\in A} d(x,y)$ *and* $d(\cdot,\cdot) = \frac{1}{n^2}\sum_{y,z\in A} d(y,z)$. *Then* $\psi(\boldsymbol{S}) = \boldsymbol{D}$, $\varphi(\boldsymbol{D}) = \boldsymbol{S}$, *and* $\varphi(\psi(\kappa)), \kappa \in \boldsymbol{S}$ *and* $\psi(\varphi(d)), d \in \boldsymbol{D}$ *are identity transformations.*

Remark 1 The $K \to \mathcal{D}$ transformation (1) is the matrix form of (2). The matrix form of (3) is

$$K = -H\mathcal{D}H + \Sigma J, \tag{4}$$

where $J = \frac{1}{n}\mathbf{1}\cdot\mathbf{1}^T$ and $H = I - J$ is the *centering matrix*.

3 Kernel, proximity and distance properties

3.1 Adjacency matrix based kernels and measures

Let us consider several kernels on graphs based on the weighted adjacency matrix W of a graph.

Katz kernel The *Katz kernel* [28] (also referred to as walk proximity [14] and von Neumann[3] diffusion kernel [37, 38]) is defined[4] as follows:

$$K^{\mathrm{Katz}}(\alpha) = \sum_{k=0}^{\infty}(\alpha W)^k = [I - \alpha W]^{-1},$$

with $0 < \alpha < (\rho(W))^{-1}$, where $\rho(W)$ is the spectral radius of W.

It is easy to see that $[I - \alpha W]$ is an M-matrix[5], i.e., a matrix of the form $A = qI - B$, where $B = (b_{ij})$ with $b_{ij} \geq 0$ for all $1 \leq i, j \leq n$, while q exceeds the maximum of the moduli of the eigenvalues of B (in the present case, $q = 1$). Thus, $[I - \alpha W]$ is a symmetric M-matrix, i.e., a Stieltjes matrix. Consequently, $[I - \alpha W]$ is positive definite and so is $K^{\mathrm{Katz}}(\alpha) = [I - \alpha W]^{-1}$. Thus, by Schoenberg's theorem, K^{Katz} can be transformed by (1) into a matrix of squared Euclidean distances.

Moreover, the Katz kernel has the following properties:
If $[I - \alpha W]$ is row diagonally dominant, i.e., $|1 - \alpha w_{ii}| \geq \alpha\sum_{j\neq i}|w_{ij}|$ for all $i \in V(G)$ (by the finiteness of the underlying space, one can always choose α small enough such that this inequality becomes valid) then

- $K^{\mathrm{Katz}}(\alpha)$ satisfies the triangle inequality for proximities (see Corollary 6.2.5 in [29]), therefore, transformation (2) provides a distance on $V(G)$;
- $K^{\mathrm{Katz}}(\alpha)$ satisfies egocentrism (i.e., *entrywise* diagonal dominance; see also Metzler's property in [29]).

Thus, in the case of row diagonal dominance of $[I - \alpha W]$, the Katz kernel is a non-normalized proximity.

[3] M. Saerens [36] has remarked that a more suitable name could be *Neumann diffusion kernel*, referring to the *Neumann series* $\sum_{k=0}^{\infty} T^k$ (where T is an operator) named after Carl Gottfried Neumann, while a connection of that to John von Neumann is not obvious (the concept of von Neumann kernel in group theory is essentially different).

[4] In fact, L. Katz considered $\sum_{k=1}^{\infty}(\alpha W)^k$.

[5] For the properties of M-matrices, we refer to [29].

Communicability kernel The *communicability kernel* [23, 20, 21] is defined as follows:

$$K^{\mathrm{comm}}(t) = \exp(tW) = \sum_{k=0}^{\infty} \frac{t^k}{k!} W^k.$$

(We shall use letter "t" whenever some notion of time can be attached to the kernel parameter; otherwise, we shall keep using letter "α".) It is an instance of symmetric exponential diffusion kernels [31]. Since K^{comm} is positive semidefinite, by Schoenberg's theorem, it can be transformed by (1) into a matrix of squared Euclidean distances. However, this does not imply that K^{comm} is a proximity.

In fact, it is easy to verify that for the graph G with adjacency matrix

$$W = \begin{pmatrix} 0 & 2 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix}, \tag{5}$$

$K^{\mathrm{comm}}(1)$ violates the triangle inequality for proximities on the triple of vertices $(1, \mathbf{2}, 3)$ (the "x" element of the inequality is given in bold). On the other hand, $K^{\mathrm{comm}}(t) \to I$ as $t \to 0$, which implies that $K^{\mathrm{comm}}(t)$ with a sufficiently small t is a [non-normalized] proximity.

Note that the graph corresponding to (5) is a weighted path 1–2–3–4, and immediate intuition suggests the inequality $d(1,2) < d(1,3) < d(1,4)$ for a distance on its vertices. However, $K^{\mathrm{comm}}(3)$ induces a Euclidean distance for which $d(1,3) > d(1,4)$. For $K^{\mathrm{comm}}(4.5)$ we even have $d(1,2) > d(1,4)$. However, $K^{\mathrm{comm}}(t)$ with a small enough positive t satisfies the common intuition.

By the way, the Katz kernel behaves similarly: when $\alpha > 0$ is sufficiently small, it holds that $d(1,2) < d(1,3) < d(1,4)$, but for $\alpha > 0.375$, we have $d(1,3) > d(1,4)$. Moreover, if $0.38795 < \alpha < (\rho(W))^{-1}$, then $d(1,2) > d(1,4)$ is true.

Double-factorial similarity The *double-factorial similarity* [22] is defined as follows:

$$K^{\mathrm{df}}(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!!} W^k.$$

As distinct from the communicability measure, K^{df} is not generally a kernel. Say, for the graph with weighted adjacency matrix (5), $K^{\mathrm{df}}(1)$ has two negative eigenvalues. Therefore K^{df} does not generally induce a set of points in $\mathbb{R}^n$, nor does it induce a natural Euclidean distance on $V(G)$.

Furthermore, in this example, matrix $\mathcal{D}$ obtained from $K^{\mathrm{df}}(1)$ by (1) has negative entries. Therefore, by Corollary 1, the function $\kappa(x,y) = K^{\mathrm{df}}_{xy}(1),\ x, y \in V(G)$ is not a proximity.

However, as well as $K^{\mathrm{comm}}(t)$, $K^{\mathrm{df}}(t) \to I$ as $t \to 0$. Consequently, all eigenvalues of $K^{\mathrm{df}}(t)$ converge to 1, and hence, $K^{\mathrm{df}}(t)$ with a sufficiently small

positive t satisfies the triangle inequality for proximities. Thus, $K^{\mathrm{df}}(t)$ with a small enough positive t is a kernel and a [non-normalized] proximity.

3.2 Laplacian based kernels and measures

Heat kernel The *heat kernel* is a symmetric exponential diffusion kernel [31] defined as follows:

$$K^{\mathrm{heat}}(t) = \exp(-tL) = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} L^k,$$

where L is the ordinary Laplacian matrix of G.

$K^{\mathrm{heat}}(t)$ is positive-definite for all values of t, and hence, it is a kernel. Then, by Schoenberg's theorem, K^{heat} induces a Euclidean distance on $V(G)$. For our example (5), this distance for all $t > 0$ obeys the intuitive inequality $d(1,2) < d(1,3) < d(1,4)$.

On the other hand, K^{heat} is not generally a proximity. E.g., for the example (5), $K^{\mathrm{heat}}(t)$ violates the triangle inequality for proximities on the triple of vertices $(1, \mathbf{2}, 3)$ whenever $t > 0.431$. As well as for the communicability kernel, $K^{\mathrm{heat}}(t)$ with a small enough t is a proximity. Moreover, it is an 1-proximity, as L has row sums 0, while $L^0 = I$ has row sums 1. Thus, the 1-normalization condition is satisfied for any $t > 0$.

Normalized heat kernel The *normalized heat kernel* is defined as follows:

$$K^{\text{n-heat}}(t) = \exp(-t\mathcal{L}) = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} \mathcal{L}^k,$$

where $\mathcal{L} = D^{-1/2} L D^{-1/2}$ is the normalized Laplacian, D being the degree matrix of G [15].

For this kernel, the main conclusions are the same as for the standard heat kernel. For the example (5), $K^{\mathrm{heat}}(t)$ violates the triangle inequality for proximities on the triple of vertices $(1, \mathbf{2}, 3)$ when $t > 1.497$. It is curious to observe that the triangle inequality of the example (5) is violated starting with a larger value of t in comparison with the case of the standard heat kernel. An important distinction is that generally, $\mathcal{L}$ has nonzero row sums. As a result, $K^{\text{n-heat}}$ does not satisfy the normalization condition, and even for small t, $K^{\text{n-heat}}$ is a non-normalized proximity.

Regularized Laplacian kernel The *regularized Laplacian kernel*, or *forest kernel* is defined [11] as follows:

$$K^{\mathrm{regL}}(t) = [I + tL]^{-1},$$

where $t > 0$.

As was shown in [12, 14], the regularized Laplacian kernel is a 1-proximity and a row stochastic matrix. Since $[I + tL]$ is positive definite, so is $[I + tL]^{-1}$, and by Schoenberg's theorem, K^{regL} induces a Euclidean distance on $V(G)$.

For the example (5), the induced distances corresponding to K^{regL} always satisfy $d(1,2) < d(1,3) < d(1,4)$. Regarding the other properties of K^{regL}, we refer to [12, 3].

It is the first encountered example of similarity measure that satisfies the both distance and proximity properties for all values of the kernel parameter.

Absorption kernel The *absorption kernel* [27] is defined as follows:

$$K^{\mathrm{absorp}}(t) = [tA + L]^{-1}, \quad t > 0,$$

where $A = \mathrm{Diag}(\boldsymbol{a})$ and $\boldsymbol{a} = (a_1, \ldots, a_n)^T$ is called the *vector of absorption rates* and has positive components. As $K^{\mathrm{absorp}}(t^{-1}) = t(A + tL)^{-1}$, this kernel is actually a generalization of the previous one.

Since $[tA + L]$ is positive definite, Schoenberg's theorem attaches a matrix of squared Euclidean distances to $K^{\mathrm{absorp}}(t)$.

$[tA + L]$ is a row diagonally dominant Stieltjes matrix, hence, by Corollary 6.2.5 in [29] we conclude that K^{absorp} satisfies the triangle inequality for proximities, i.e., K^{absorp} is a proximity (but not generally a Σ-proximity).

3.3 Markov matrix based kernels and measures

Personalized PageRank *Personalized PageRank* (PPR) similarity measure is defined as follows:

$$K^{\mathrm{PPR}}(\alpha) = [I - \alpha P]^{-1},$$

where $P = D^{-1}W$ is a row stochastic (Markov) matrix, D is the degree matrix of G, and $0 < \alpha < 1$, which corresponds to the standard random walk on the graph.

In general, $K^{\mathrm{PPR}}(\alpha)$ is not symmetric, so it is not positive semidefinite, nor is it a proximity.

Moreover, the functions $d(x, y)$ obtained from K^{PPR} by transformation[6]

$$d(x, y) = \frac{1}{2}(\kappa(x, x) + \kappa(y, y) - \kappa(x, y) - \kappa(y, x)) \tag{6}$$

need not generally be distances. Say, for

$$W = \begin{pmatrix} 0 & 2 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 2 & 0 \end{pmatrix} \tag{7}$$

[6] If K is symmetric, then (6) coincides with (2).

with $K^{\mathrm{PPR}}(\alpha)$, one has $d(1,3)+d(3,4)<d(1,4)$ whenever $\alpha>0.9515$.

K^{PPR} has only positive eigenvalues. However, its symmetrized counterpart $\frac{1}{2}(K^{\mathrm{PPR}}+(K^{\mathrm{PPR}})^T)$ may have a negative eigenvalue (say, with $\alpha\geq 0.984$ for (5) or with $\alpha\geq 0.98$ for (7)). Thus, it need not be positive semidefinite and, consequently, by Theorem 1, $\mathcal{D}$ obtained from it by (1) (or from K^{PPR} by (6)) is not generally a matrix of squared Euclidean distances.

K^{PPR} satisfies the normalization condition. For a small enough α, it can be transformed (as well as K^{comm} and K^{df}) into a distance matrix using (6).

On the other hand, one can slightly modify Personalized PageRank so it becomes a proximity. Rewrite K^{PPR} as follows:

$$[I-\alpha D^{-1}W]^{-1}=[D-\alpha W]^{-1}D.$$

Then, consider

Modified Personalized PageRank

$$K^{\mathrm{modifPPR}}(\alpha)=[I-\alpha D^{-1}W]^{-1}D^{-1}=[D-\alpha W]^{-1},\quad 0<\alpha<1,$$

which becomes a non-normalized proximity by Corollary 6.2.5 in [29]. In particular, the triangle inequality becomes

$$\frac{K^{\mathrm{PPR}}_{ii}(\alpha)}{d_i}-\frac{K^{\mathrm{PPR}}_{ji}(\alpha)}{d_i}-\frac{K^{\mathrm{PPR}}_{ik}(\alpha)}{d_k}+\frac{K^{\mathrm{PPR}}_{jk}(\alpha)}{d_k}\geq 0,$$

which looks like an interesting inequality for Personalized PageRank. Due to symmetry, $K^{\mathrm{modifPPR}}_{ij}=K^{\mathrm{modifPPR}}_{ji}$, and we obtain an independent proof of the following identity for Personalized PageRank [2]:

$$\frac{K^{\mathrm{PPR}}_{ij}(\alpha)}{d_j}=\frac{K^{\mathrm{PPR}}_{ji}(\alpha)}{d_i}.$$

Note that replacing the Laplacian matrix $L=D-W$ with $D-\alpha W$ is a kind of alternative regularization of L. Being diagonally dominant,

$$D-\alpha W=\bar{d}I-(\bar{d}I-D+\alpha W) \tag{8}$$

(where $\bar{d}$ is the maximum degree of the vertices of G) is a Stieltjes matrix. Consequently, $D-\alpha W$ is positive definite and so is $K^{\mathrm{modifPPR}}(\alpha)=[D-\alpha W]^{-1}$. Thus, by Schoenberg's theorem, K^{modifPPR} can be transformed by (1) into a matrix of squared Euclidean distances.

We note that Personalized PageRank can be generalized by using non-homogeneous restart [4], which will lead to the discrete-time analog of the absorption kernel. However, curiously enough, the discrete-time version has a smaller number of proximity-distance properties than the continuous-time version.

PageRank heat similarity measure *PageRank heat similarity measure* [16] is defined as follows:

$$K^{\mathrm{heatPPR}}(t) = \exp(-t(I - P)).$$

Basically, the properties of this measure are similar to those of the standard Personalized PageRank. Say, for the example (7) with K^{heatPPR}, one has $d(1,2)+d(2,3) < d(1,3)$ whenever $t > 1.45$.

3.4 Logarithmic similarity measures and transitional properties

Given a strictly positive similarity measure $s(x, y)$, the function $\kappa(x, y) = \ln s(x, y)$ is the corresponding *logarithmic similarity*.

Using Theorem 2 it can be verified [8] that whenever $S = (s_{ij}) = (s(i, j))$ produces a strictly positive *transitional measure* on G (i.e., $s_{ij}\, s_{jk} \le s_{ik}\, s_{jj}$ for all vertices i, j, and k, while $s_{ij}\, s_{jk} = s_{ik}\, s_{jj}$ if and only if every path from i to k visits j), we haves that the logarithmic similarity $\kappa(x, y) = \ln s(x, y)$ produces a *cutpoint additive distance*, viz., a distance that satisfies $d(i, j) + d(j, k) = d(i, k)$ iff every path from i to k visits j:

$$d(i,j) = \tfrac{1}{2}(\kappa(i,i) + \kappa(j,j) - \kappa(i,j) - \kappa(j,i)) = \ln \sqrt{\frac{s(i,i)s(j,j)}{s(i,j)s(j,i)}}. \qquad (9)$$

In the case of digraphs, five transitional measures were indicated in [8], namely, *connection reliability*, *path accessibility* with a sufficiently small parameter, *walk accessibility*, and two versions of *forest accessibility*; the undirected counterparts of the two latter measures were studied in [10] and [9], respectively.

Proposition 1 *K^{absorp}, K^{PPR}, and K^{modifPPR} produce transitional measures.*

Proof. For $K^{\mathrm{absorp}}(t) = [tA + L]^{-1}$, let $h = \max_i\{a_i t + d_i - w_{ii}\}$, where d_i is the degree of vertex i. Then $K^{\mathrm{absorp}}(t) = [hI - (hI - tA - D + W)]^{-1} = [I - W']^{-1}h^{-1}$, where $W' = h^{-1}(hI - tA - D + W)$ is nonnegative with row sums less than 1. Hence, $K^{\mathrm{absorp}}(t)$ is positively proportional to the matrix $[I - W']^{-1}$ of walk weights of the graph with weighted adjacency matrix W'.

Similarly, by (8), $K^{\mathrm{modifPPR}}(\alpha) = [D - \alpha W]^{-1} = [I - W'']^{-1}\bar{d}^{-1}$, where $W'' = \bar{d}^{-1}(\bar{d}I - D + \alpha W)$ is nonnegative with row sums less than 1. Consequently, $K^{\mathrm{modifPPR}}(\alpha)$ is proportional to the matrix of walk weights of the graph whose weighted adjacency matrix is W''.

Finally, $K^{\mathrm{PPR}}(\alpha)$ is the matrix of walk weights of the digraph with weighted adjacency matrix αP.

Since by [8, Theorem 6], any finite matrix of walk weights of a weighted digraph produces a transitional measure, so do K^{absorp}, K^{PPR}, and K^{modifPPR}. □

Thus, as by Proposition 1 and the results of [8], K^{Katz}, K^{regL}, K^{absorp}, K^{PPR}, and K^{modifPPR} produce transitional measures, we have that the corresponding *logarithmic* dissimilarities (9) are cutpoint additive distances.

Furthermore, if $S = (s_{ij}) = (s(i,j))$ produces a strictly positive transitional measure on G, then, obviously, $\kappa(x,y) = \ln s(x,y)$ satisfies $\kappa(y,x) + \kappa(x,z) - \kappa(y,z) \leq \kappa(x,x)$, which coincides[7] with the triangle inequality for proximities whenever $s(x,y)$ is symmetric. Therefore, as K^{Katz}, K^{regL}, K^{absorp}, and K^{modifPPR} are symmetric, we obtain that the corresponding logarithmic similarities $\kappa(x,y) = \ln s(x,y)$ are proximities.

K^{PPR} is not generally symmetric, however, it can be observed that $\widetilde{K}^{\mathrm{PPR}}$ such that $\widetilde{K}^{\mathrm{PPR}}_{ij} = \sqrt{K^{\mathrm{PPR}}_{ij} K^{\mathrm{PPR}}_{ji}}$ is symmetric and produces the same *logarithmic* distance (9) as K^{PPR}. Hence, the logarithmic similarity $\kappa(x,y) = \ln \widetilde{K}^{\mathrm{PPR}}_{xy}$ is a proximity.

At the same time, the above logarithmic similarities are not kernels, as the corresponding matrices have negative eigenvalues.

This implies that being a proximity is not a stronger property than being a kernel. By Corollary 1, the square root of the distance induced by a proximity is also a distance. However, this square rooted distance need not generally be Euclidean, thus, Theorem 1 is not sufficient to conclude that the initial proximity is a kernel.

It can be verified that all logarithmic measures corresponding to the similarity measures under study preserve the natural order of distances $d(1,2) < d(1,3) < d(1,4)$ for the example (5).

4 Numerical comparison of similarity measures in the context of unsupervised learning

Here we compare various kernels, proximities, and generally similarity measures in the context of unsupervised learning method – spectral clustering (for background on spectral clustering see, e.g., [19, 42]). We test them on random undirected graphs that are built according to the stochastic block model.

More precisely, each graph $G = (V, E)$ has the following structure: it consists of two clusters $V = C_1 \cup C_2$ with the intracluster edge density p_{in} and the intercluster density p_{out}, i.e.

$$p_{\mathrm{in}} = P\{(i,j) \in E \mid i,j \in C_1\} = P\{(i,j) \in E \mid i,j \in C_2\},$$
$$p_{\mathrm{out}} = P\{(i,j) \in E \mid i \in C_1, j \in C_2\}.$$

We introduce the following reference classification:

$$\mathrm{cls}_{\mathrm{true}}[i] = k \text{ if } i \in C_k.$$

Given a similarity measure (matrix) K, which is computed using one of the basic graph matrices (W, L or P), we apply the spectral clustering algorithm to separate the graph into m clusters, $m = 2$ in our case. The algorithm we use is similar to the one proposed in [19]. Let us recap it here for completeness:

[7] On various alternative versions of the triangle inequality, we refer to [17].

find normalized eigenvectors of K that correspond to its largest m eigenvalues and put them into columns of matrix X; flip signs of column of X in a way that elements with maximum absolute values in each column are positive; run K-means algorithm on rows of X with m output clusters and place the result into the array cls.

We will measure the difference between two clusterings with m clusters on a set of n nodes by the following function:

$$\mathcal{E}(\mathrm{cls}_1, \mathrm{cls}_2) = 1 - \frac{1}{n} \max_{\sigma \in S_m} |\{i \in \{1, .., n\} | \sigma(\mathrm{cls}_1[i]) = \mathrm{cls}_2[i]\}|.$$

Here S_m is the group of all permutations on the set $\{1, 2, .., m\}$. This function corresponds to the minimum relative classification error that can be achieved by renumbering the clusters. Its computation is equivalent to solving the assignment problem of size m.

Since the transformation from W to K^{Katz}, K^{comm}, K^{df} is monotonic for eigenvalues and does not affect eigenvectors, these similarity measures all lead to the same result by spectral clustering. The similarity measures K^{heat} and K^{regL} are in the same sense equivalent to $-L = W - D$. $K^{\mathrm{n-heat}}$ is equivalent to $-\mathcal{L} = -D^{-1/2} L D^{-1/2}$ and to $D^{-1/2} W D^{-1/2}$. K^{PPR} and K^{heatPPR} are equivalent to $P = D^{-1} W$.

Hence, it is meaningful to test the clustering procedure on W, $-L$, $-\mathcal{L}$ and P. Looking ahead, we say that for the unbalanced case that we tested the typical error for W and $-L$ was about 0.4 that was much more than for other similarity measures. Hence, we included results only on P, $-\mathcal{L}$ and also added the results of spectral clustering algorithm from scikit-learn Python library, which is based on left eigenvectors of P.

The logarithmic transformation, however, changes both eigenvalues and eigenvectors. Hence, it is interesting to test it for different similarity measures. Since they all depend on some parameters, we minimize the error over the parameter space for each graph and than average it over the set of graph realizations.

4.1 Balanced model

We tested the unsupervised learning algorithms on 100 graph realizations of 200 nodes stochastic block model with two clusters of 100 nodes each, intracluster density $p_{\mathrm{in}} = 0.1$ and intercluster density $p_{\mathrm{out}} = 0.02$. Errors, minimized over the parameter space and averaged over 100 graph realizations are shown in Figure 1. The black thin bars correspond to the 95% confidence intervals. Spectral sklearn corresponds to the spectral clustering algorithm from scikit-learn Python library. Spectral P and Spectral NL correspond to the spectral clustering algorithm with $P = D^{-1} W$ and $I - \mathcal{L} = D^{-1/2} W D^{-1/2}$. The others correspond to the spectral clustering algorithm with logarithmic transformations of corresponding similarity measures. Black errorbars correspond to the 95% confidence interval.

We observe that all the tested methods provide roughly the same error that is around 0.01%. This is the manifestation of the fact that in the balanced case clustering is relatively easy.

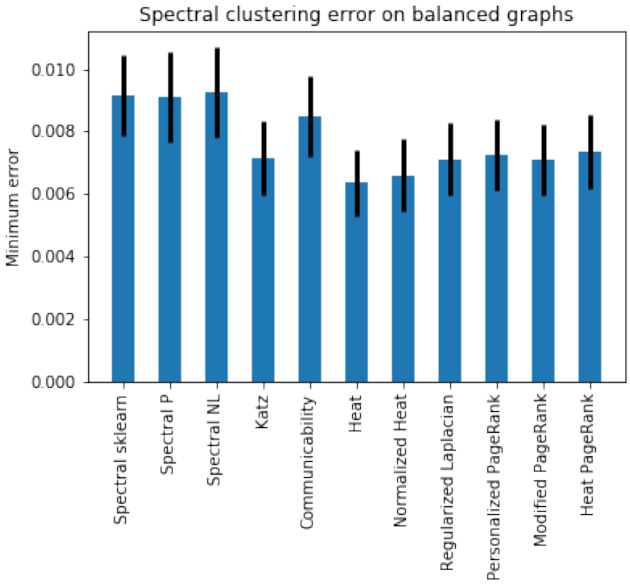

Fig. 1. Averaged minimum error for the balanced model.

4.2 Unbalanced model

We also tested the algorithms on 1000 graphs of 200 nodes with two clusters of 50 and 150 nodes and the same edge densities $p_{\text{in}} = 0.1$ and $p_{\text{out}} = 0.02$. As expected, clustering unbalanced classes is more challenging.

Here we observe significant difference between results obtained with different similarity measures. Katz, communicability and normalized heat log measures lead to best results in this case.

We note that some other aspects of the comparative behavior of several kernels on graphs in clustering tasks have been studied in [26, 30, 40].

Acknowledgements

The work of KA and DR was supported by the joint Bell Labs Inria ADR "Network Science" and by UCA-JEDI Idex Grant "HGRAPHS", and the work of PC was supported by the Russian Science Foundation (project no.16-11-00063 granted to IRE RAS). This is an author-edited copy of the article published in Proceedings of the 14th Workshop on Algorithms and Models for the Web Graph (WAW 2017), Springer LNCS v.10519.

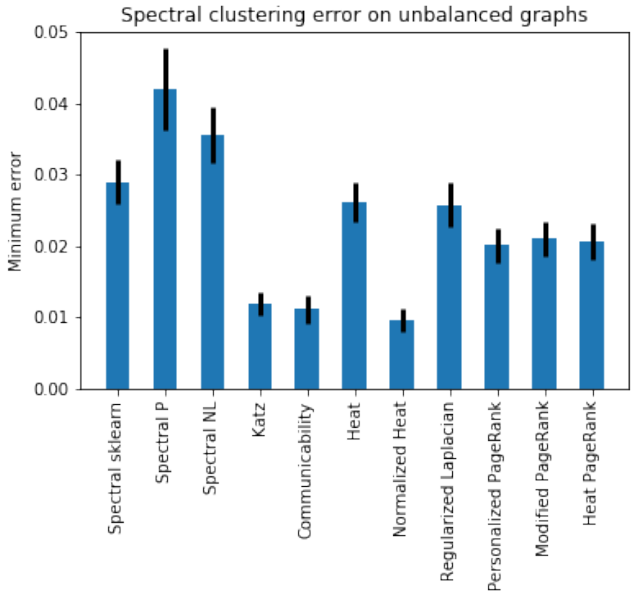

Fig. 2. Averaged minimum error for the unbalanced model.

References

1. Avrachenkov, K., Mishenin, A., Gonçalves, P. and Sokol, M., 2012. Generalized optimization framework for graph-based semi-supervised learning. In Proceedings of the 2012 SIAM International Conference on Data Mining (pp. 966-974).
2. Avrachenkov, K., Gonçalves, P. and Sokol, M., 2013. On the choice of kernel and labelled data in semi-supervised learning methods. In Proccedings of WAW 2013 (pp. 56-67).
3. Avrachenkov, K., Chebotarev, P. and Mishenin, A., 2017. Semi-supervised learning with regularized Laplacian. Optimization Methods and Software, 32(2), pp. 222-236.
4. Avrachenkov, K., van der Hofstad, R. and Sokol, M., 2014. Personalized PageRank with node-dependent restart. In *Proceedings of International Workshop on Algorithms and Models for the Web-Graph* (pp. 23-33).
5. Backstrom, L. and Leskovec, J., 2011. Supervised random walks: predicting and recommending links in social networks. Proceedings of ACM WSDM 2011, pp. 635-644.
6. Boley, D., Ranjan, G. and Zhang, Z.L., 2011. Commute times for a directed graph using an asymmetric Laplacian. Linear Algebra and its Applications, 435(2), pp. 224-242.
7. Chapelle, O., Schölkopf, B. and Zien, A. 2006. *Semi-Supervised Learning*, MIT Press.
8. Chebotarev, P., 2011. The graph bottleneck identity. Advances in Applied Mathematics, 47(3), pp. 403-413.

9. Chebotarev, P., 2011. A Class of graph-geodetic distances generalizing the shortest-path and the resistance distances. Discrete Applied Mathematics, 47(3), pp. 403-413.
10. Chebotarev, P., 2012. The walk distances in graphs, Discrete Applied Mathematics, 160(10-11), pp. 1484-1500.
11. Chebotarev, P.Yu. and Shamis, E.V., 1995. On the proximity measure for graph vertices provided by the inverse Laplacian characteristic matrix. In *Abstracts of the conference "Linear Algebra and its Application"*, 10-12 June 1995, The Institute of Mathematics and its Applications, in conjunction with the Manchester Center for Computational Mathematics, Manchester, UK (pp. 6-7). URL `http://www.ma.man.ac.uk/~higham/laa95/abstracts.ps`
12. Chebotarev, P.Yu. and Shamis, E.V., 1997. The matrix-forest theorem and measuring relations in small social groups. Autom. Remote Control, 58(9), pp. 1505-1514.
13. Chebotarev, P.Yu. and Shamis, E.V., 1998. On a duality between metrics and Σ-proximities. Autom. Remote Control, 59(4), pp. 608-612.
14. Chebotarev, P.Yu. and Shamis, E.V., 1998. On proximity measures for graph vertices. Autom. Remote Control, 59(10), pp. 1443-1459.
15. Chung, F., 1997. *Spectral graph theory.* American Math. Soc.
16. Chung, F., 2007. The heat kernel as the pagerank of a graph. Proceedings of the National Academy of Sciences, 104(50), pp. 19735-19740.
17. Deza, M. and Chebotarev, P., 2011. Protometrics. arXiv preprint arXiv:1112.4829.
18. Dhillon, I. S., Fan, J. and Guan, Y., 2001. Efficient clustering of very large document collections. Data mining for scientific and engineering applications, 2, pp. 357-381.
19. Dhillon, I. S., Guan, Y. and Kulis, B., 2004. Kernel k-means: spectral clustering and normalized cuts. Proceedings of ACM KDD 2004, pp. 551-556.
20. Estrada, E. and Hatano, N., 2007. Statistical-mechanical approach to subgraph centrality in complex networks. Chem. Phys. Lett., 439, pp. 247-251.
21. Estrada, E. and Hatano, N., 2008. Communicability in complex networks. Physical Review E, 77(3), 036111.
22. Estrada, E. and Silver, G., 2017. Accounting for the role of long walks on networks via a new matrix function. Journal of Mathematical Analysis and Applications, 449, pp. 1581-1600.
23. Fouss, F., Yen L., Pirotte, A., and Saerens, M., 2006. An experimental investigation of graph kernels on a collaborative recommendation task. In *Proceedings of the Sixth International Conference on Data Mining (ICDM'06).* IEEE. (pp. 863-868).
24. Fouss, F., Saerens, M., and Shimbo, M., 2016. *Algorithms and Models for Network Data and Link Analysis.* Cambridge University Press.
25. Horn, R.A. and Johnson, C.R., 2013. Matrix Analysis (2nd Edition). Cambridge University Press.
26. Ivashkin,V. and Chebotarev P., 2017. Do logarithmic proximity measures outperform plain ones in graph clustering? In V.A. Kalyagin et al., eds. *Models, Algorithms, and Technologies for Network Analysis*, Proceedings in Mathematics & Statistics, Vol. 197, Springer, In press.
27. Jacobsen, K. and Tien, J., 2016. A generalized inverse for graphs with absorption. ArXiv preprint arXiv:1611.02233.
28. Katz, L., 1953. A new status index derived from sociometric analysis. Psychometrika, 18(1), pp. 39-43.
29. Kirkland, S.J. and Neumann, M., 2012. *Group Inverses of M-matrices and Their Applications.* CRC Press.

30. Kivimäki, I., Shimbo, M. and Saerens, M., 2014. Developments in the theory of randomized shortest paths with a comparison of graph node distances. Physica A: Statistical Mechanics and its Applications, 393, pp. 600616.
31. Kondor, R.I. and Lafferty, J., 2002. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of ICML* (pp. 315-322).
32. Lenart, C., 1998. A generalized distance in graphs and centered partitions. SIAM Journal on Discrete Mathematics, 11(2), pp. 293-304.
33. Liben-Nowell, D. and Kleinberg, J., 2007. The link-prediction problem for social networks. Journal of the Association for Information Science and Technology, 58(7), pp. 1019-1031.
34. Schoenberg, I.J., 1935. Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert". Ann. Math. 36(3), pp. 724-732.
35. Schoenberg, I.J., 1938. Metric spaces and positive definite functions. Transactions of the American Mathematical Society, 44(3), pp. 522-536.
36. Saerens, M. Personal communication.
37. Kandola, J., Shawe-Taylor, J. and Cristianini, N., 2002. Learning semantic similarity. In *Neural Information Processing Systems 15 (NIPS 15)*. MIT Press.
38. Shawe-Taylor, J. and Cristianini, N., 2004. *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press.
39. Smola, A.J. and Kondor, R., 2003. Kernels and regularization on graphs. In *Learning Theory and Kernel Machines* (pp. 144-158).
40. Sommer, F., Fouss, F. and Saerens, M., 2016. Comparison of graph node distances on clustering tasks, *Lecture Notes in Computer Science*, LNCS 9886, Springer, pp. 192201.
41. Vishwanathan, S.V.N., Schraudolph, N.N., Kondor, R. and Borgwardt, K.M., 2010. Graph kernels. Journal of Machine Learning Research, 11(Apr), pp. 1201-1242.
42. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing, v.17(4), pp.395-416 (2007).
43. Zhou, D., Schölkopf, B. and Hofmann, T., 2004. Semi-supervised learning on directed graphs. In *Proceeedings of NIPS* (pp. 1633-1640).
44. Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., 2001. An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks*, Vol. 12, No. 2 (pp. 181-202).