ARTICLE IN PRESS

European Journal of Combinatorics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

European Journal of Combinatorics

journal homepage: www.elsevier.com/locate/ejc

Similarities on graphs: Kernels versus proximity measures

Konstantin Avrachenkov [a], Pavel Chebotarev [b,c,d], Dmytro Rubanov [a]

[a] *Inria Sophia Antipolis, France*
[b] *Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russia*
[c] *Kotelnikov Institute of Radioengineering and Electronics of RAS, Moscow, Russia*
[d] *Moscow Institute of Physics and Technology, Moscow, Russia*

ARTICLE INFO

Article history:
Available online xxxx

ABSTRACT

We analytically study proximity and distance properties of various kernels and similarity measures on graphs. This helps to understand the mathematical nature of such measures and can potentially be useful for recommending the adoption of specific similarity measures in data analysis.

1. Introduction

Until the 1960s, mathematicians studied only one distance for graph vertices, the shortest path distance [6]. In 1967, Gerald Sharpe proposed the electric distance [42]; then it was rediscovered several times. For some collection of graph distances, we refer to [19, Chapter 15].

Distances[1] are treated as dissimilarity measures. In contrast, similarity measures are maximized when every distance is equal to zero, i.e., when two arguments of the function coincide. At the same time, there is a close relation between distances and certain classes of similarity measures.

One of such classes consists of functions defined with the help of kernels on graphs, i.e., positive semidefinite matrices with indices corresponding to the nodes. Every kernel is the Gram matrix of some set of vectors in a Euclidean space, and Schoenberg's theorem [40,41] shows how it can be transformed into the matrix of Euclidean distances between these vectors.

E-mail addresses: k.avrachenkov@inria.fr (K. Avrachenkov), upi@ipu.ru (P. Chebotarev), dmytro.rubanov@inria.fr (D. Rubanov).

[1] In this paper, the term distance is used as a synonym of metric, i.e., every distance satisfies nonnegativity, symmetry, the identity of indiscernibles, and the triangle inequality.

https://doi.org/10.1016/j.ejc.2018.02.002
0195-6698/

Another class consists of proximity measures characterized by the triangle inequality for proximities. Every proximity measure $\kappa(x, y)$ generates a distance [15] by means of the $d(x, y) = \frac{1}{2}(\kappa(x, x) + \kappa(y, y)) - \kappa(x, y)$ transformation, and for some subclasses of proximity measures, the inverse transformation can also be specified.

Furthermore, it turns out that many similarity measures are transitional measures [9,11], in particular, they satisfy the inequality $s_{ij}\, s_{jk} \leq s_{ik}\, s_{jj}$ and so they can be transformed into proximities by means of the logarithmic transformation.

Distances and similarity measures on graphs are widely used in data analysis, especially, in graph-based supervised, semi-supervised, and unsupervised machine learning, see, e.g., [1,2,4,5,7,21,25,26,35–37,47] and references therein. In a number of studies including [28,33,45], comparative ability of various measures to detect communities and predict links has been explored. However, in such studies, the authors do not focus on the mathematical properties of the measures under consideration.

The purpose of this paper is to start filling this gap. We consider a number of well-known and recently proposed similarity measures on graphs (including weighted graphs) defined in terms of one of the following basic matrices: the weighted adjacency matrix, the Laplacian matrix, and the (stochastic) Markov matrix. We explore their basic mathematical properties, in particular, we find out whether they belong to the classes of kernels or proximities and study the properties of distances related to them. This helps to reveal the nature of such measures and can be considered as a step towards a mathematical theory of similarity / dissimilarity measures on graphs.

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

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

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 $\kappa : V(G) \times V(G) \to \mathbb{R}$ 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 [2,20,25,36].

A kernel on graph is a graph similarity measure that has an inner product representation. All the 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 [27].

We note that following [34,44] we prefer to write *kernel on graph* rather than *graph kernel*, as the notion of "graph kernel" refers to a kernel between graphs [46].

A *proximity measure* (or simply *proximity*) [15] on a set A is a function $\kappa : A \times A \to \mathbb{R}$ 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 $x, y \in A$. Consequently, if $\kappa(x, y)$ is a proximity, then $-\kappa(x, y)$ is a protometric [18,19].

Furthermore, any Σ-proximity has the *egocentrism* property: $\kappa(x, x) > \kappa(x, y)$ for any distinct $x, y \in A$ [15]. If $\kappa(x, y)$ is represented by a matrix $K = (K_{xy}) = (\kappa(x, y))$, then egocentrism of $\kappa(x, y)$ amounts to the *strict entrywise diagonal dominance* of K.

Please cite this article in press as: K. Avrachenkov, et al., Similarities on graphs: Kernels versus proximity measures, European Journal of Combinatorics (2018), https://doi.org/10.1016/j.ejc.2018.02.002.