Physica A 393 (2014) 600–616

Contents lists available at ScienceDirect

Physica A

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

Developments in the theory of randomized shortest paths with a comparison of graph node distances

Ilkka Kivimäki [a,*], Masashi Shimbo [b], Marco Saerens [a]

[a] ICTEAM Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[b] Nara Institute of Science and Technology, Ikoma, Japan

HIGHLIGHTS

- A generalized graph node distance measure is derived based on Helmholtz free energy.
- A new algorithm for computing randomized shortest path dissimilarities is presented.
- A comparison of generalized graph node distances is performed.
- The comparison is based on visualization, clustering and classification tasks.
- The free energy distance provides competitive results in the comparative experiments.

ARTICLE INFO

Article history:
Received 25 March 2013
Received in revised form 13 August 2013
Available online 12 September 2013

Keywords:
Graph node distances
Free energy
Randomized shortest paths
Shortest path distance
Commute time and resistance distance
Clustering

ABSTRACT

There have lately been several suggestions for parametrized distances on a graph that generalize the shortest path distance and the commute time or resistance distance. The need for developing such distances has risen from the observation that the above-mentioned common distances in many situations fail to take into account the global structure of the graph. In this article, we develop the theory of one family of graph node distances, known as the randomized shortest path dissimilarity, which has its foundation in statistical physics. We show that the randomized shortest path dissimilarity can be easily computed in closed form for all pairs of nodes of a graph. Moreover, we come up with a new definition of a distance measure that we call the free energy distance. The free energy distance can be seen as an upgrade of the randomized shortest path dissimilarity as it defines a metric, in addition to which it satisfies the graph-geodetic property. The derivation and computation of the free energy distance are also straightforward. We then make a comparison between a set of generalized distances that interpolate between the shortest path distance and the commute time, or resistance distance. This comparison focuses on the applicability of the distances in graph node clustering and classification. The comparison, in general, shows that the parametrized distances perform well in the tasks. In particular, we see that the results obtained with the free energy distance are among the best in all the experiments.

1. Introduction

Defining distances and similarities between nodes of a graph based on its structure has become an essential task in the analysis of network data [1–9]. In the simplest case, a binary network can be presented as an adjacency matrix or adjacency

* Corresponding author. Tel.: +32 10478388.
E-mail address: ilkka.kivimaki@uclouvain.be (I. Kivimäki).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.09.016

list which can be difficult to interpret. Acquiring meaningful information from such data requires sophisticated methods which often need to be chosen based on the context. Being able to measure the distance between the nodes of a network in a meaningful way of course provides a fundamental way of interpreting the network. With the information of distances between the nodes, one can apply traditional multivariate statistical or machine learning methods for analyzing the data.

The most common ways of defining a distance on a graph are to consider either the lengths of the shortest paths between nodes, leading to the definition of the *shortest path (SP) distance*, or the expected lengths of random walks on the graph, which can be used to derive the *commute time (CT) distance* [10]. The CT distance is known to equal the *resistance distance* [11,12] up to a constant factor [13]. In this paper, we examine generalized distances on graphs that interpolate, depending on a parameter, between the shortest path distance and the commute time or resistance distance.

The paper contains several separate contributions: First, we develop the theory of one generalized distance, the *randomized shortest path (RSP) dissimilarity* [14,15]. We derive a new algorithm for computing it for all pairs of nodes of a graph in closed form, and thus much more efficiently than before. We then derive another generalized distance from the RSP framework based on the Helmholtz free energy between two states of a thermodynamic system. We show that this *free energy (FE) distance* actually coincides with the *potential distance*, proposed in recent literature in a more ad hoc manner [16]. However, our new derivation gives a nice theoretical background for this distance. Finally, we make a comparison of the behavior and performance of different generalized graph node distances. The comparisons are conducted by observing the relative differences of distances between nodes in small example graphs and by examining the performance of the different distance measures in clustering and classification tasks.

The paper is structured as follows: In Section 2, we define the terms and notation used in the paper. In our framework, we consider graphs where the edges are assigned weights and costs, which can be independent of each other. In Section 3, we recall the definitions of the common distances on graphs. We also present a surprising result related to the generalization of the commute time distance considering costs, namely that the distance based on costs equals the commute time distance, up to a constant factor. In Section 4, we revisit the definition of the RSP dissimilarity [14,15]. We then derive the closed form algorithm, mentioned above, for computing it, and then formulate the definition of the FE distance. In Section 5, we present other parametrized distances on graphs interpolating between the SP and CT distances that have been defined in recent literature. Section 6 contains the comparison of the RSP dissimilarity, the FE distance and the generalized distances defined in Section 5. Finally, Section 7 sums up the content of the article.

2. Terminology and notation

We first go through the terminology and notation used in this paper. We denote by $G = (V, E)$ a graph G consisting of a node set $V = \{1, 2, \ldots, n\}$ and an edge set $E = \{(i, j)\}$. Nodes i and j such that $(i, j) \in E$ are called *adjacent* or *connected*. Each graph can be represented as an adjacency matrix $\mathbf{A}$, where the elements a_{ij} are called *affinities*, or *weights*, interchangeably. For unweighted graphs $a_{ij} = 1$ if $(i, j) \in E$, for weighted graphs $a_{ij} > 0$ if $(i, j) \in E$ and in both cases $a_{ij} = 0$ if $(i, j) \notin E$. The affinities can be interpreted as representing the degree of similarity between connected nodes. A *path*, or *walk*, interchangeably, on the graph G is a sequence of nodes $\wp = (i_0, \ldots, i_T)$, where $T \geq 0$ and $(i_\tau, i_{\tau+1}) \in E$ for all $\tau = 0, \ldots, T - 1$. The *length* of the path, or walk, $\wp$, is then T. Note that throughout this article we include zero-length paths $(i),\ i \in V$ in the definition of a path, although in some contexts it may be more appropriate to disallow this by setting $T \geq 1$ in the definition. Moreover, we define *absorbing*, or *hitting* paths as paths which contain the terminal node only once. Thus a path $\wp$ is an absorbing path if $\wp = (i_0, \ldots, i_T)$, where $i_T \neq i_\tau$ for all $\tau = 0, \ldots, T - 1$.

In addition to affinities, the edges of a graph can be assigned *costs*, c_{ij}, such that $0 < c_{ij} < \infty$ if $(i, j) \in E$. The cost of a path $\wp$ is the sum of the costs along the path[1] $\widetilde{c}(\wp) = \sum_{(i,j)\in\wp} c_{ij}$. In principle, we do not define costs for unconnected pairs of nodes, but when making matrix computations, we assign the corresponding matrix elements a very large number (compared to other costs). When there is no natural cost assigned to the edges, a common convention is to define the costs as reciprocals of the affinities $c_{ij} = 1/a_{ij}$. This applies both for unweighted and weighted graphs. This way the edge weights and costs are analogous to conductance and resistance, respectively, in an electric network. In the experiments, in Section 6 of this paper, we always use this conversion for determining costs from affinities. However, in the theory that we present in Sections 3–4, we consider that the costs can also be assigned independently of the affinities, allowing a more general setting. This can be useful in many applications because links can often have a two-sided nature, on one hand based on the structure of the graph and on the other hand based on internal features of the edges. One such example can be a toll road network, where the affinities represent the proximities of places and the costs represent toll costs of traversing a road. This interpretation is especially useful in graph analysis based on a probabilistic framework, wherein the emphasis of this paper also lies. Experiments that take advantage of the possible independence between affinities and costs are left for further work.

We denote by $\mathbf{e}$ the $n \times 1$ vector whose each element is 1. For an $n \times n$ square matrix $\mathbf{A}$, let $\mathbf{Diag}(\mathbf{A})$ denote the $n \times n$ diagonal matrix whose diagonal elements are the diagonal elements of $\mathbf{A}$ and by $\mathbf{diag}(\mathbf{A})$ the $n \times 1$ vector of the diagonal elements of $\mathbf{A}$. Likewise, for an $n \times 1$ vector $\mathbf{v}$, $\mathbf{Diag}(\mathbf{v})$ denotes the $n \times n$ diagonal matrix containing the elements of vector $\mathbf{v}$ on its diagonal. We use $\exp(\mathbf{A})$ and $\log(\mathbf{A})$ to denote the elementwise exponential and logarithm, respectively; these should

[1] Throughout the article we will use the tilde ($\sim$) to differentiate quantities related to paths from quantities related to edges.