A Bag-of-Paths Framework for Network Data Analysis

Kevin Françoisse[a], Ilkka Kivimäki[a,c], Amin Mantrach[d],
Fabrice Rossi[e], Marco Saerens[a,b]

[a] *Université catholique de Louvain, Belgium*
[b] *Université Libre de Bruxelles, Belgium*
[c] *Aalto University, Department of Computer Science, Helsinki, Finland*
[d] *Yahoo! Research, Sunnyvale, California, USA*
[e] *Université Paris 1 Panthéon-Sorbonne, France*

Abstract

This work develops a generic framework, called the bag-of-paths (BoP), for link and network data analysis. The central idea is to assign a probability distribution on the set of all paths in a network. More precisely, a Gibbs-Boltzmann distribution is defined over a bag of paths in a network, that is, on a representation that considers all paths independently. We show that, under this distribution, the probability of drawing a path connecting two nodes can easily be computed in closed form by simple matrix inversion. This probability captures a notion of relatedness, or more precisely accessibility, between nodes of the graph: two nodes are considered as highly related when they are connected by many, preferably low-cost, paths. As an application, two families of distances between nodes are derived from the BoP probabilities. Interestingly, the second distance family interpolates between the shortest-path distance and the commute-cost distance. In addition, it extends the Bellman-Ford formula for computing the shortest-path distance in order to integrate sub-optimal paths (exploration) by simply replacing the minimum operator by the soft minimum operator. Experimental results on semi-supervised classification tasks show that both of the new distance families are competitive with other state-of-the-art approaches. In addition to the distance measures studied in this paper, the bag-of-paths framework enables straightforward computation of many other relevant network measures.

Keywords: Network science, link analysis, distance and similarity on a graph, shortest-path distance, resistance distance, commute-time distance, semi-supervised classification.

1. Introduction

1.1. General introduction

Network and link analysis is a highly studied field, subject of much recent work in various areas of science: applied mathematics, computer science, social science, physics, chemistry, pattern recognition, applied statistics, data mining and machine learning, to name a few [4, 21, 32, 60, 66, 78, 93, 103, 109]. Within this context, one key issue is the proper quantification of the structural relatedness between nodes of a network by taking both direct and indirect connections into account [68]. This problem is faced in all disciplines involving networks in

various types of problems such as link prediction, community detection, node classification, and network visualization to name a few popular ones.

The main contribution of this paper is in presenting in detail the **bag-of-paths** (BoP) framework and defining relatedness as well as distance measures between nodes from this framework. The BoP builds on and extends previous work dedicated to the exploratory analysis of network data [58, 57, 72, 113]. The introduced distances are constructed to capture the global structure of the graph by using paths on the graph as a building block. In addition to relatedness/distance measures, various other quantities of interest can be derived within the probabilistic BoP framework in a principled way, such as betweenness measures quantifying to which extent a node is in between two sets of nodes [64], extensions of the modularity criterion for, e.g., community detection [28], measures capturing the criticality of the nodes or robustness of the network [65], graph cuts based on BoP probabilities, and so on.

1.2. The bag-of-paths framework

More precisely, we assume given a weighted directed, strongly connected, graph or network G where a cost is associated to each edge. Within this context, we consider a bag containing all the possible (either absorbing or non-absorbing – see later for details) paths[1] between pairs of nodes in G. In a first step, following [2, 72, 87, 113], a probability distribution on this countable set of paths can be defined by minimizing the total expected cost between all pairs of nodes while fixing the total relative entropy spread in the graph. This results in a Gibbs-Boltzmann distribution, depending on a temperature parameter T, on the set of paths such that long (high-cost) paths have a low probability of being sampled from the bag, while short (low-cost) paths have a high probability of being sampled.

In this probabilistic framework, the **BoP probabilities**, $\mathrm{P}(s = i, e = j)$, that a sampled path has node i as its starting node and node j as its ending node can easily be computed in closed form by a simple $n \times n$ matrix inversion, where n is the number of nodes in the graph. These BoP probabilities play a crucial role in our framework for that they capture the *relatedness* (in terms of accessibility) between each pair of nodes (i, j): the BoP probability will be high when the two nodes are connected by many, short, paths. In summary, the BoP framework has several interesting properties:

- It has a clear, intuitive, interpretation.
- The temperature parameter T allows to monitor randomness by controlling the balance between exploitation and exploration.
- The introduction of independent costs results in a large degree of customization of the model, according to the problem requirements: some paths could be penalized because they visit undesirable nodes having adverse features.
- The framework is rich. Many useful quantities of interest can be defined according to the BoP probabilistic framework: similarity and distance

[1] These are also called walks in the literature.

measures, betweenness measures, etc. This is discussed extensively in the sequel.

- The quantities of interest are easy to compute.

It, however, also suffers from a drawback: the different quantities are computed by solving a system of linear equations, or by matrix inversion. More precisely, the distance between a particular node and all the other nodes can be computed by solving a system of n linear equations, while all pairwise distances can be computed at once by inverting an $n \times n$ square matrix. This results in $\mathcal{O}(n^3)$ computational complexity. Even more importantly, the matrix of distances necessitates $\mathcal{O}(n^2)$ storage, although this can be alleviated by using, e.g., incomplete matrix factorization techniques.

This means that the different quantities can only be computed reasonably on small to medium size graphs (containing a few tens of thousand nodes). However, in specific applications like classification or extraction of top eigenvectors, we can avoid computing explicitly the matrix inversion (see PageRank and the power method [63], or large scale semi-supervised classification on graphs [71]). In addition, it is also possible to restrict the set of paths to "efficient paths", that is, paths that do not backtrack (always getting further from the starting node), and compute efficiently the distances from the starting node by a recurrence formula, as proposed in transportation theory [29].

1.3. Defining node distances from the bag-of-paths framework

The paper first introduces the BoP framework in detail and derives the BoP probabilities from this framework. Thereafter, two families of distances between nodes are defined, and are coined the **surprisal distance** and the **potential distance**. Both distance measures satisfy the triangle inequality, and thus satisfy the axioms of a metric. Moreover, the potential distance has the interesting property of generalizing the shortest-path and the commute-cost distances by computing an intermediate distance, depending on a temperature parameter T. When T is close to zero, the distance reduces to the standard shortest-path distance (emphasizing exploitation) while for $T \to \infty$, it reduces to the commute-cost distance (focusing on exploration). The commute-cost distance is closely related to the resistance distance and the commute-time distance [35, 59], as the three functions are proportional to each other [12, 58].

This is of primary interest as it has been shown that both the shortest-path distance and the resistance distance suffer from some significant flaws. While relevant in many applications, the shortest-path distance cannot always be considered as a good candidate distance in network data. Indeed, this measure only depends on the shortest paths and thus does not integrate the "degree of connectivity" between the two nodes. In many applications, for a constant shortest-path distance, nodes connected by many indirect paths should be considered as "closer" than nodes connected by only a few paths. This is especially relevant when considering relatedness of nodes based on communication, movement, etc, in a network which do not always happen optimally, nor completely randomly. Moreover, the shortest path distance suffers from another flaw: when computing the distance from a given node, the shortest-path distance usually provides many ties, that is, many nodes with the same distance, especially in weighted, undirected, graphs. Breaking these ties can be done by considering other properties, like the amount of connectivity.

While the shortest-path distance fails to take the whole structure of the graph into account, it has also been shown that the resistance and the commute-time distances converge to a useless value, only depending on the degrees of the two nodes, when the size of the graph increases (the random walker is getting "lost in space" because the Markov chain mixes too fast; see [106, 107]). Moreover, the resistance distance, which is proportional to the commute-time distance, assumes completely random movements or communication in the network, which is also unrealistic.

In short, shortest paths do not integrate the amount of connectivity between the two nodes whereas random walks quickly loose the notion of proximity to the initial node when the graph becomes larger [106, 107].

There is therefore a need for introducing distances interpolating between the shortest-path distance and the resistance distance, thus hopefully avoiding the drawbacks appearing at the ends of the spectrum. These quantities capture the notion of relative *accessibility* between nodes, a combination of both proximity in the network and amount of connectivity (see the next Section 2 for a short survey).

Furthermore, and interestingly, a simple local recurrence expression, extending the Bellman-Ford formula for computing the potential distances from one node of interest to all the other nodes is also derived. It relies on the use of the so-called soft minimum operator [24] instead of the usual minimum. Finally, our experiments show that these distance families provide competitive results in semi-supervised learning.

1.4. Contributions and organization of the paper

Thus, in summary, this work has several contributions:

- It introduces a well-founded bag-of-paths framework capturing the global structure of the graph by using network paths as a building block.
- It is shown that the bag-of-hitting-paths probabilities can easily be computed in closed form. This fundamental quantity defines an intuitive relatedness measure between nodes.
- It defines two families of distances from the bag-of-paths probabilities, capturing the structural dissimilarity between the nodes in terms of relative accessibility. The distances between all pairs of nodes can be computed conveniently by inverting an $n \times n$ matrix.
- It is shown that one of these distance measures has some interesting properties; for instance it is cutpoint additive and it interpolates between the shortest-path distance and the resistance distance (up to a scaling factor).
- The framework is extended to the case where non-uniform priors are defined on the nodes.
- We prove that this distance generalizes the Bellman-Ford formula computing shortest-path distances, by simply replacing the min operator by the softmin operator.
- The distances obtain promising empirical results in semi-supervised classification tasks when compared to other, kernel-based, methods.

Section 2 develops related work and introduces the necessary background and notation. Section 3 introduces the BoP framework, defines BoP probabilities and shows how it can be computed in closed form. Section 4 extends the framework to hitting, or absorbing, paths. In Section 5, the two families of distances as well as their properties are derived. Section 6 generalizes the framework to non-uniform priors on the nodes. An experimental study of the BoP framework with application to semi-supervised classification is presented in Section 7. Concluding remarks and extensions are discussed in Section 8.

2. Related work, background, and notation

2.1. Related work

This work is related to similarity measures on graphs for which some background is presented in this section, largely inspired by the surveys appearing in [34, 58, 72, 112, 113]. The presented BoP framework also has applications in semi-supervised classification, on which our experimental section will focus on in Section 7. A short survey related to this problem can be found in subsection 7.1.

Similarity measures on a graph determine to what extent two nodes in a graph resemble each other, either based on the information contained in the node attributes or based on the graph structure. In this work, only measures based on the graph structure will be investigated. Structural similarity measures can be categorized into two groups: local and global [68]. On the one hand, local similarity measures between nodes consider the direct links from a node to the other nodes as features and use these features in various way to provide similarities. Examples include the cosine coefficient [31] and the standard correlation [109]. On the other hand, global similarity measures consider the whole graph structure to compute similarities.

Certainly the most popular and useful distance between nodes of a graph is the shortest-path distance. However, as discussed in the introduction, it is not always relevant for quantifying the similarity of nodes in a network.

Alternatively, similarity measures can be based on random walk models on the graph, seen as a Markov chain. As an example, the commute-time (CT) kernel has been introduced in [35, 88] as the Moore-Penrose pseudoinverse, $\mathbf{L}^+$, of the Laplacian matrix. The CT kernel was inspired by the work of Klein & Randic [59] and Chandra et al. [12]. More precisely, Klein & Randic [59] suggested to use the effective resistance between two nodes as a meaningful distance measure, called the resistance distance. Chandra et al. [12] then showed that the resistance distance equals the commute-time distance (but also the commute-cost distance), up to a constant factor. The CT distance is defined as the average number of steps that a random walker, starting in a given node, will take before entering another node for the first time (this is called the average first-passage time [79]) and going back to the initial node.

It was then shown [88] that the elements of $\mathbf{L}^+$ are inner products of the node vectors in the Euclidean space where these node vectors are exactly separated by the square root of the CT distance. The square root of the CT distance is therein called the Euclidean CT distance. The relationships between the Laplacian matrix and the *commute-cost* distance (the expected *cost* (and not steps as for the CT) of reaching a destination node from a starting node and

going back to the starting node) were studied in [35]. Finally, an electrical interpretation of the elements of $\mathbf{L}^+$ can be found in [112]. However, we saw in the introduction that these random-walk based distances suffer from some drawbacks (e.g., the so-called "lost in space" problem, [106, 107]).

Sarkar et al. [89] suggested a fast method for computing truncated commute-time neighbors. At the same time, several authors defined an embedding that preserves the commute-time distance with applications in various fields such as clustering [115], collaborative filtering [35, 10], dimensionality reduction of manifolds [42] and image segmentation [84].

Instead of taking the pseudoinverse of the Laplacian matrix, a simple regularization leads to a kernel called the regularized commute-time kernel [48], with a nice interpretation in terms of the matrix-forest theorem [18, 19]. Ito et al. [48], further propose the modified regularized Laplacian kernel by introducing another parameter controlling the importance of nodes. This modified regularized Laplacian kernel is also closely related to a graph regularization framework introduced by Zhou & Scholkopf in [120], extended to directed graphs in [119].

The adjacency-based exponential diffusion kernel, a member of the more general exponential diffusion kernel family introduced by Kondor & Lafferty [62] and investigated in [36, 34], as well as the Neumann diffusion kernel introduced in [90], are both based on power series of the adjacency matrix. A meaningful alternative to the adjacency-based exponential diffusion kernel, called the Laplacian exponential diffusion kernel [62, 94], is a diffusion model that substitutes the adjacency matrix with minus the Laplacian matrix, and is closely related to the heat kernel [22]. It is also a member of the more general exponential diffusion kernel family, computing the matrix exponential of a matrix reflecting the local structure of the graph (e.g., the adjacency matrix or the Laplacian matrix, see [62, 94] and [36, 34] for empirical comparisons). Note that the adjacency-based exponential diffusion kernel is also known as the communicability measure in the physics community [33, 32].

Random walk with restart kernels, inspired by the PageRank algorithm and adapted to provide relative similarities between nodes, appeared relatively recently in [83, 80, 105]. Nadler et al. [75, 76] and Pons et al. [81, 82] suggested a distance measure between nodes of a graph based on a diffusion process, called the diffusion distance. The Markov diffusion kernel has been derived from this distance measure in [34] and [114]. The natural embedding induced by the diffusion distance was called diffusion map by Nadler et al. [75, 76] and is related to correspondence analysis [114].

More recently, Mantrach et al. [72], inspired by [2, 6] and subsequently by [87], introduced a link-based covariance measure between nodes of a weighted directed graph, called the sum-over-paths covariance. They consider, in a similar manner as in this paper, a Gibbs-Boltzmann distribution on the set of paths such that high-cost paths occur with low probability whereas low-cost paths occur with a high probability. Their main goal was to define a well-founded covariance/correlation measure between nodes based on the following property: two nodes are considered as highly correlated if they often co-occur together on the same – preferably short – paths. In other words, two nodes obtain a high correlation index when they are likely to appear on the same walk. Note that a related co-betweenness measure between nodes has been introduced in [61]. The present work re-interprets the idea introduced in [72] from a bag-of-paths point of view, extends the results, and focuses on the study of *accessibility*

measures between pairs of nodes, integrating both proximity in the graph and high connectivity between the nodes.

Indeed, as both the shortest-path distance and the resistance distance show some issues, there were several attempts to define *families of distances* interpolating between the shortest-path and more "global" distances, such as the resistance distance. In this context, inspired by [2, 6, 87], a parametrized family of dissimilarity measures, called the randomized shortest-path (RSP) dissimilarity, reducing to the shortest-path distance at one end of the parameter range, and to the resistance distance (up to a constant scaling factor) at the other end, was proposed in [113] and extended in [58], for the efficient computation of the entire matrix of dissimilarities. The RSP dissimilarity between two nodes i, j is simply the expected cost for visiting a node j for the first time when starting from a node i, under the assumption that paths between i and j are chosen thanks to a Gibbs-Boltzmann distribution. The calculation of this quantity requires the computation of a forward and a backward variable, as in hidden Markov models [37]. Similar ideas appeared at the same time in [15, 16, 17], based on considering the co-occurences of nodes in forests of a graph or walks on the graph, and in [44, 3], based on a generalization of the effective resistance in electric circuits. These two last families are metrics while the RSP dissimilarity does not satisfy the triangle inequality. The potential and the surprisal distances introduced in this work fall under the same catalogue of distance families. See also [58, 41, 40] for other, closely related, formulations of families of distances based on free energy and network flows.

2.2. Some background and notation

We now introduce the necessary notation for the bag-of-paths (BoP) framework. First, note that, in the sequel, column vectors are written in bold lowercase while matrices are in bold uppercase.

Consider a weighted directed graph or network, $G = (\mathcal{V}, \mathcal{E})$, assumed strongly connected, with a set $\mathcal{V}$ of n nodes (or vertices) and a set $\mathcal{E}$ of edges (or arcs, links). An edge between node i and node j is denoted by $i \to j$ or (i, j). Furthermore, it is assumed that we are given an adjacency matrix $\mathbf{A}$ with elements $a_{ij} \geq 0$ quantifying in some way the affinity between node i and node j. When $a_{ij} > 0$, node i and node j are said to be adjacent, that is, connected by an edge. Conversely, $a_{ij} = 0$ means that i and j are not connected. We further assume that there are no self-loops, that is, the $a_{ii} = 0$.

From this adjacency matrix, a standard random walk on the graph is defined in the usual way. The transition probabilities associated to each node are simply proportional to the affinities and then normalized:

$$p_{ij}^{\text{ref}} = \frac{a_{ij}}{\sum_{j'=1}^{n} a_{ij'}} \tag{1}$$

Note that these transition probabilities will be used as reference probabilities later; hence the superscript "ref". The matrix $\mathbf{P}^{\text{ref}}$, containing elements p_{ij}^{ref}, is stochastic and called the transition matrix of the natural or reference random walk on the graph.

In addition, we assume that a transition cost, $c_{ij} \geq 0$, is associated to each edge $i \to j$ of the graph G. If there is no edge between i and j, the cost is assumed to take an infinite value, $c_{ij} = \infty$. For consistency, $c_{ij} = \infty$ if and only

if $a_{ij} = 0$. The cost matrix $\mathbf{C}$ is the matrix containing the immediate costs c_{ij} as elements. We will assume that at least one element of $\mathbf{C}$ is strictly positive. A path $\wp$ is a finite sequence of jumps to adjacent nodes on G (including loops), initiated from a starting node $s = i$, and stopping in an ending node $e = j$. The *total cost* of a path $\wp$ is simply the sum of the local costs c_{ij} along $\wp$, while the *length* of a path is the number of steps, or jumps, needed for following that path.

The costs are set independently of the adjacency matrix; they quantify the cost of a transition, depending on the problem at hand. They can, e.g., be defined according to some properties, or features, of the nodes or the edges in order to bias the probability distribution of choosing a path. In the case of a social network, we may, for instance, want to bias the paths in favor of domain experts. In that case, the cost of jumping to a node could be set proportional to the degree of expertise of the corresponding person. Therefore, walks visiting a large proportion of persons with a low degree of expertise would be penalized versus walks visiting persons with a high degree. Another example aims to favor hub-avoiding paths penalizing paths visiting hubs. Then, the cost can be simply set to the degree of the node. If there are no natural costs associated to edges and there is no reason to bias the paths with respect to some features, costs are simply set equal to 1 (paths are penalized by their length) or set to $c_{ij} = 1/a_{ij}$ (the elements of the adjacency matrix can then be considered as conductances and the costs as resistances).

3. The basic bag-of-paths framework

Recall that the bag-of-paths (BoP) model will be based on the probability that a path drawn from a "bag of paths" has nodes i and j as its starting and ending nodes, respectively. This probability distribution then serves as a building block for various extensions.

In this section, the bag-of-paths framework is introduced by first considering bounded paths and then paths of arbitrary length. For simplicity, we discuss non-hitting (or non-absorbing) paths first and then develop the more interesting bag-of-hitting-paths framework in the next section 4.

3.1. Sampling bounded paths according to a Gibbs-Boltzmann distribution

Following [2, 72, 87, 113], the present section describes how the probability distribution on the set of paths is assigned. In order to make the presentation more rigorous, we will first have to consider paths of *bounded length* t. Later, we will extend the results for paths with arbitrary length. Let us first choose two nodes, a starting node i and an ending node j and define the set of paths (including cycles) of length t from i to j as $\mathcal{P}_{ij}(t) = \{\wp_{ij}(t)\}$. Thus, $\mathcal{P}_{ij}(t)$ contains all the paths $\wp_{ij}(t)$ allowing to reach node j from node i in *exactly* t *steps*.

Let us further denote as $\tilde{c}(\wp_{ij}(t))$ the total cost associated to path $\wp_{ij}(t)$. Here, we assume that $\wp_{ij}(t)$ is a valid path from node i to node j, that is, it consists of a sequence of nodes $(k_0 = i) \to k_1 \to k_2 \to \cdots \to (k_t = j)$ where $c_{k_{\tau-1}k_\tau} < \infty$ for all $\tau \in [1, t]$. As already mentioned, we assume that the total cost associated to a path is additive, i.e. $\tilde{c}(\wp_{ij}(t)) = \sum_{\tau=1}^{t} c_{k_{\tau-1}k_\tau}$. Then, let us

define the set of all t-length paths through the graph between all pairs of nodes as $\mathcal{P}(t) = \cup_{i,j=1}^{n} \mathcal{P}_{ij}(t)$.

Finally, the set of all bounded paths *up to length* t is denoted by $\mathcal{P}(\leq t) = \cup_{\tau=0}^{t} \mathcal{P}(\tau)$. Note that, by convention, for $i = j$ and $t = 0$, zero-length paths are allowed with zero associated cost. Other types of paths will be introduced later; a summary of the mathematical notation appears in Table 1.

Now, we consider a probability distribution on this finite set $\mathcal{P}(\leq t)$, representing the probability of drawing a path $\wp \in \mathcal{P}(\leq t)$ from a bag containing all paths up to length t. We search for the distribution of paths $\mathrm{P}(\wp)$ minimizing the expected total cost-to-go, $\mathbb{E}[\tilde{c}(\wp)]$, among all the distributions having a fixed relative entropy J_0 with respect to a reference distribution, here the natural random walk on the graph (see Equation (1)). This choice naturally defines a probability distribution on the set of paths of maximal length t such that high-cost paths occur with a low probability while short paths occur with a high probability. In other words, we are seeking for path probabilities, $\mathrm{P}(\wp)$, $\wp \in \mathcal{P}(\leq t)$, minimizing the expected total cost subject to a constant relative entropy constraint[2]:

$$\left| \begin{array}{ll} \underset{\{\mathrm{P}(\wp)\},\, \wp \in \mathcal{P}(\leq t)}{\text{minimize}} & \displaystyle\sum_{\wp \in \mathcal{P}(\leq t)} \mathrm{P}(\wp)\tilde{c}(\wp) \\ \text{subject to} & \sum_{\wp \in \mathcal{P}(\leq t)} \mathrm{P}(\wp) \log(\mathrm{P}(\wp)/\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp)) = J_0 \\ & \sum_{\wp \in \mathcal{P}(\leq t)} \mathrm{P}(\wp) = 1 \end{array} \right. \tag{2}$$

where $J_0 > 0$ is provided a priori by the user, according to the desired degree of randomness and $\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp)$ represents the probability of following the path $\wp$ when walking according to the reference transition probabilities p_{ij}^{ref} of the natural random walk on G (see Equation (1)).

More precisely, we define $\tilde{\pi}^{\mathrm{ref}}(\wp) = \prod_{\tau=1}^{t} p_{k_{\tau-1}k_\tau}^{\mathrm{ref}}$, that is, the product of the transition probabilities along path $\wp$ – the likelihood of the path when the starting and ending nodes are known. Now, if we assume a uniform (non-uniform priors are considered in Section 4), independent, a priori probability, $1/n$, for choosing the starting and the ending node, then we set $\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp) = \tilde{\pi}^{\mathrm{ref}}(\wp) / \sum_{\wp' \in \mathcal{P}(\leq t)} \tilde{\pi}^{\mathrm{ref}}(\wp')$, which ensures that the reference probability is properly normalized[3].

The problem (2) can be solved by introducing the following Lagrange function

$$\mathscr{L} = \sum_{\wp \in \mathcal{P}(\leq t)} \mathrm{P}(\wp)\tilde{c}(\wp) + \lambda \left[\sum_{\wp \in \mathcal{P}(\leq t)} \mathrm{P}(\wp) \log \frac{\mathrm{P}(\wp)}{\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp)} - J_0 \right] + \mu \left[\sum_{\wp \in \mathcal{P}(\leq t)} \mathrm{P}(\wp) - 1 \right] \tag{3}$$

and optimizing over the set of path probabilities $\{\mathrm{P}(\wp)\}_{\wp \in \mathcal{P}(\leq t)}$. As could be expected (the problem is similar to a maximum entropy problem [50, 55]), set-

[2] In theory, non-negativity constraints should be added, but this is not necessary as the resulting probabilities are automatically non-negative.

[3] We will see later that the path likelihoods $\tilde{\pi}^{\mathrm{ref}}(\wp)$ are already properly normalized in the case of hitting, or absorbing, paths: $\sum_{\wp \in \mathcal{P}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp) = 1$. See next section and Appendix A. On the contrary, for regular paths as considered in this section, it can be shown that $\sum_{\wp \in \mathcal{P}(\leq t)} \tilde{\pi}^{\mathrm{ref}}(\wp) = (t+1)n$ (we thank Dr. Guillaume Guex for deriving and letting us know this result).

ting its partial derivative with respect to $\mathrm{P}(\wp)$ to zero and solving the equation yields a **Gibbs-Boltzmann probability distribution** on the set of paths up to length t [72],

$$\mathrm{P}(\wp) = \frac{\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp'\in\mathcal{P}(\leq t)}\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \tag{4}$$

where the Lagrange parameter λ plays the role of a temperature T and $\theta = 1/\lambda$ is the inverse temperature.

Thus, as desired, short paths $\wp$ (having a low cost $\tilde{c}(\wp)$) are favored in that they have a large probability of being followed. Moreover, from Equation (4), we clearly observe that when $\theta \to 0$, the path probabilities reduce to the probabilities generated by the natural random walk on the graph (characterized by the transition probabilities p_{ij}^{ref} as defined in Equation (1)). In this case, $J_0 \to 0$ as well. But when θ is large, the probability distribution defined by Equation (4) is biased towards low-cost paths (the most likely paths are the shortest ones).

Note that, in the sequel, it will be assumed that the user provides the value of the parameter θ instead of J_0, with $\theta > 0$. Also notice that the model could be derived thanks to a maximum entropy principle instead [50, 55].

3.2. The bag-of-paths probabilities

Let us now derive an important quantity from Equation (4), namely the probability of drawing a path starting in some node $s = i$ and ending in some other node $e = j$ from the bag of paths. These quantities will be called the (bounded) **bag-of-paths** (BoP) **probabilities**. For paths up to length t this is provided by

$$\begin{aligned}\mathrm{P}^{(\leq t)}(s=i,e=j) &= \frac{\sum_{\wp\in\mathcal{P}_{ij}(\leq t)}\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp'\in\mathcal{P}(\leq t)}\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \\ &= \frac{\sum_{\wp\in\mathcal{P}_{ij}(\leq t)}\tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp'\in\mathcal{P}(\leq t)}\tilde{\pi}^{\mathrm{ref}}(\wp')\exp[-\theta\tilde{c}(\wp')]}\end{aligned} \tag{5}$$

where we recall that $\mathcal{P}_{ij}(\leq t)$ is the set of paths up to length t connecting node i to node j. From (4), this quantity simply computes the probability mass of drawing a path connecting i to j. The paths in $\mathcal{P}_{ij}(\leq t)$ can contain loops and could visit nodes i and j several times during the trajectory[4].

3.2.1. Computation of the bag-of-paths probabilities for bounded paths

The analytical expression allowing to compute the quantity defined by Equation (5) will be derived in this subsection. Then, in the following subsection, its

[4]Note that another interesting class of paths, the hitting, or absorbing, paths – allowing only one single visit to the ending node j – will be considered in the next section 4.

$\wp$	a particular path visiting nodes $k_0, k_1, \ldots, k_t$
$\mathrm{P}(\wp)$	the probability of drawing path $\wp$
$\mathcal{P}_{ij}(t)$	set of paths connecting i to j in exactly t steps
$\mathcal{P}_{ij}(\leq t)$	set of paths connecting i to j in at most t steps
$\mathcal{P}(\leq t) = \cup_{i,j=1}^{n} \mathcal{P}_{ij}(\leq t)$	set of all paths of at most t steps
$\mathcal{P}_{ij}$	set of paths of arbitrary length connecting i to j
$\mathcal{P} = \cup_{i,j=1}^{n} \mathcal{P}_{ij}$	set of all paths of arbitrary length
$\mathbf{P}^{\mathrm{ref}}$	transition probability matrix with elements p_{ij}^{ref}
$\mathbf{C}$	cost matrix with elements c_{ij}
$\tilde{\pi}^{\mathrm{ref}}(\wp) = \prod_{\tau=1}^{t} p_{k_{\tau-1}k_\tau}^{\mathrm{ref}}$	likelihood of following path $\wp$ according to p_{ij}^{ref}
$\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp) = \tilde{\pi}^{\mathrm{ref}}(\wp) / \sum_{\wp' \in \mathcal{P}} \tilde{\pi}^{\mathrm{ref}}(\wp')$	normalized likelihood of following path $\wp$
$\tilde{c}(\wp)$	total cumulated cost when following path $\wp$

Table 1: Summary of notation for the enumeration of paths in a graph G.

definition will be extended to the set of paths of arbitrary length (unbounded paths) by taking the limit $t \to \infty$.

We start from the cost matrix, $\mathbf{C}$, from which we build a new matrix, $\mathbf{W}$, as

$$\mathbf{W} = \mathbf{P}^{\mathrm{ref}} \circ \exp[-\theta \mathbf{C}] = \exp\left[-\theta \mathbf{C} + \log \mathbf{P}^{\mathrm{ref}}\right] \tag{6}$$

where $\mathbf{P}^{\mathrm{ref}}$ is the transition probability matrix[5] of the natural random walk on the graph containing the elements p_{ij}^{ref}, and the logarithm/exponential functions are taken elementwise. Moreover, $\circ$ is the elementwise (Hadamard) matrix product. Note that the matrix $\mathbf{W}$ is not symmetric in general.

Then, let us first compute the numerator of Equation (5). Because all the quantities in the exponential of Equation (5) are summed along a path, $\log \tilde{\pi}^{\mathrm{ref}}(\wp) = \sum_{\tau=1}^{t} \log p_{k_{\tau-1}k_\tau}^{\mathrm{ref}}$ and $\tilde{c}(\wp) = \sum_{\tau=1}^{t} c_{k_{\tau-1}k_\tau}$ where each link $k_{\tau-1} \to k_\tau$ lies on path $\wp$, we immediately observe that element i, j of the matrix $\mathbf{W}^\tau$ ($\mathbf{W}$ to the power τ) is $[\mathbf{W}^\tau]_{ij} = \sum_{\wp \in \mathcal{P}_{ij}(\tau)} \exp[-\theta \tilde{c}(\wp) + \log \tilde{\pi}^{\mathrm{ref}}(\wp)]$ where $\mathcal{P}_{ij}(\tau)$ is the set of paths connecting the starting node i to the ending node j in *exactly* τ steps.

Consequently, the sum in the numerator of Equation (5) is

$$\begin{aligned} \sum_{\wp \in \mathcal{P}_{ij}(\leq t)} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)] &= \sum_{\tau=0}^{t} \sum_{\wp \in \mathcal{P}_{ij}(\tau)} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)] \\ &= \sum_{\tau=0}^{t} [\mathbf{W}^\tau]_{ij} = \left[\sum_{\tau=0}^{t} \mathbf{W}^\tau\right]_{ij} = \mathbf{e}_i^{\mathrm{T}} \left(\sum_{\tau=0}^{t} \mathbf{W}^\tau\right) \mathbf{e}_j \end{aligned} \tag{7}$$

where $\mathbf{e}_i$ is a column vector full of 0's, except in position i where it contains a 1. By convention, at time step 0, the random walker appears in node i with probability one and a zero cost: $\mathbf{W}^0 = \mathbf{I}$. This means that *zero-length paths* (without any transition step) are allowed in $\mathcal{P}_{ij}(\leq t)$. If, on the contrary, we want to dismiss zero-length paths, we could redefine $\mathcal{P}_{ij}(\leq t)$ as the set of paths

[5] Do not confuse matrix $\mathbf{P}^{\mathrm{ref}}$ in bold with $\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp)$ representing the reference probability of path $\wp$. A summary of the notation appears in Table 1.

of length at least one (the summation starts at $t = 1$ instead of $t = 0$) and proceed in the same manner.

This previous Equation (7) allows to derive the analytical form of the probability of drawing a bounded path (up to length t) starting in node i and ending in j. Indeed, replacing Equation (7) in Equation (5), and recalling that $\mathcal{P}(\leq t) = \cup_{i,j=1}^{n} \mathcal{P}_{ij}(\leq t)$, we obtain

$$\mathrm{P}^{(\leq t)}(s = i, e = j) = \frac{\mathbf{e}_i^{\mathrm{T}} \left(\sum_{\tau=0}^{t} \mathbf{W}^{\tau} \right) \mathbf{e}_j}{\sum_{i,j=1}^{n} \mathbf{e}_i^{\mathrm{T}} \left(\sum_{\tau=0}^{t} \mathbf{W}^{\tau} \right) \mathbf{e}_j} = \frac{\mathbf{e}_i^{\mathrm{T}} \left(\sum_{\tau=0}^{t} \mathbf{W}^{\tau} \right) \mathbf{e}_j}{\mathbf{e}^{\mathrm{T}} \left(\sum_{\tau=0}^{t} \mathbf{W}^{\tau} \right) \mathbf{e}} \tag{8}$$

where $\mathbf{e} = [1, 1, \ldots, 1]^{\mathrm{T}}$ is a vector of 1's. But, of course, there is no a priori reason to choose a particular path length; we will therefore consider paths of arbitrary length in the next section.

3.2.2. Proceeding with paths of arbitrary length

Let us now consider the problem of computing the probability of drawing a path starting in i and ending in j from a bag containing paths of *arbitrary* length, and therefore usually containing an infinite, but countable, number of paths. Following the definition in the bounded case (Equation (5)), this quantity will be denoted as and defined by

$$\mathrm{P}(s = i, e = j) = \lim_{t \to \infty} \mathrm{P}^{(\leq t)}(s = i, e = j) = \frac{\sum_{\wp \in \mathcal{P}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)]}{\sum_{\wp' \in \mathcal{P}} \tilde{\pi}^{\mathrm{ref}}(\wp') \exp[-\theta \tilde{c}(\wp')]} \tag{9}$$

where $\mathcal{P}_{ij}$ is the set of paths (of all lengths) connecting i to j in the graph and the denominator is called the **partition function** of the bag-of-paths system,

$$\mathcal{Z} = \sum_{\wp \in \mathcal{P}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)] \tag{10}$$

As already stated before, the quantity $\mathrm{P}(s = i, e = j)$ in Equation (9) will be called the **bag-of-paths probability** of drawing a path of arbitrary length starting from node i and ending in node j. This key quantity captures a notion of *accessibility* (and thus relatedness or similarity in terms of accessibility), between nodes of G. From Equation (9), we observe that two nodes are considered as highly accessible (high probability of sampling them) when they are connected by many, preferably low-cost paths. The quantity therefore integrates the concept of connectivity (amount of paths), in addition to proximity (low-cost paths).

Now, from Equation (8), we need to compute

$$\mathrm{P}(s = i, e = j) = \lim_{t \to \infty} \mathrm{P}^{(\leq t)}(s = i, e = j) = \lim_{t \to \infty} \frac{\mathbf{e}_i^{\mathrm{T}} \left(\sum_{\tau=0}^{t} \mathbf{W}^{\tau} \right) \mathbf{e}_j}{\mathbf{e}^{\mathrm{T}} \left(\sum_{\tau=0}^{t} \mathbf{W}^{\tau} \right) \mathbf{e}} \tag{11}$$

We thus need to compute the well-known power series of $\mathbf{W}$

$$\lim_{t\to\infty}\sum_{\tau=0}^{t}\mathbf{W}^{\tau}=\sum_{t=0}^{\infty}\mathbf{W}^{t}=(\mathbf{I}-\mathbf{W})^{-1} \tag{12}$$

which converges if the spectral radius of $\mathbf{W}$ is less than 1, $\rho(\mathbf{W})<1$. Because the matrix $\mathbf{W}$ only contains non-negative elements and G is strongly connected, a sufficient condition for $\rho(\mathbf{W})<1$ is that it is substochastic [74], which is always achieved for $\theta>0$ as $c_{ij}\geq 0$ for all i,j and we assume that at least one element of $\mathbf{C}$ is strictly positive. We therefore assume a $\theta>0$.

Now, if we pose

$$\mathbf{Z}=(\mathbf{I}-\mathbf{W})^{-1} \tag{13}$$

with $\mathbf{W}$ given by Equation (6), we can pursue the computation of the numerator of Equation (11),

$$\mathbf{e}_i^{\mathrm{T}}\left(\sum_{t=0}^{\infty}\mathbf{W}^{t}\right)\mathbf{e}_j=\mathbf{e}_i^{\mathrm{T}}(\mathbf{I}-\mathbf{W})^{-1}\mathbf{e}_j=\mathbf{e}_i^{\mathrm{T}}\mathbf{Z}\mathbf{e}_j=[\mathbf{Z}]_{ij}=z_{ij} \tag{14}$$

where z_{ij} is element i,j of $\mathbf{Z}$. By analogy with Markov chain theory, $\mathbf{Z}$ is called the **fundamental matrix** [56]. Elementwise, following Equations (7) and (14), we have that

$$z_{ij}=\sum_{\wp\in\mathcal{P}_{ij}}\tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]=\left[(\mathbf{I}-\mathbf{W})^{-1}\right]_{ij} \tag{15}$$

which is actually related to the potential of a Markov chain [23, 79]. From the previous equation, z_{ij} can be interpreted as

$$\begin{aligned}z_{ij}=\sum_{t=0}^{\infty}[\mathbf{W}^{t}]_{ij}&=\delta_{ij}+p_{ij}^{\mathrm{ref}}\,e^{-\theta c_{ij}}+\sum_{k_1=1}^{n}p_{ik_1}^{\mathrm{ref}}p_{k_1j}^{\mathrm{ref}}\,e^{-\theta(c_{ik_1}+c_{k_1j})}\\&+\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}p_{ik_1}^{\mathrm{ref}}p_{k_1k_2}^{\mathrm{ref}}p_{k_2j}^{\mathrm{ref}}\,e^{-\theta c_{ik_1}}e^{-\theta c_{k_1k_2}}e^{-\theta c_{k_2j}}+\cdots\end{aligned} \tag{16}$$

For the denominator of Equation (9) and (11), we immediately find

$$\mathcal{Z}=\mathbf{e}^{\mathrm{T}}\mathbf{Z}\mathbf{e}=z_{\bullet\bullet} \tag{17}$$

where $z_{\bullet\bullet}=\sum_{i,j=1}^{n}z_{ij}$ is the value of the partition function $\mathcal{Z}$. Therefore, from Equation (11), the probability of drawing a path starting in i and ending in j in our bag-of-paths model is simply

$$\mathrm{P}(s=i,e=j)=\frac{z_{ij}}{\mathcal{Z}},\ \text{with}\ \mathbf{Z}=(\mathbf{I}-\mathbf{W})^{-1}\ \text{and}\ \mathcal{Z}=z_{\bullet\bullet} \tag{18}$$

or, in matrix form,

$$\mathbf{\Pi}=\frac{\mathbf{Z}}{z_{\bullet\bullet}},\ \text{with}\ \mathbf{Z}=(\mathbf{I}-\mathbf{W})^{-1} \tag{19}$$

where $\mathbf{\Pi}$, called the **bag-of-paths probability matrix**, contains the probabilities for each starting-ending pair of nodes. Note that this matrix is not symmetric in general; therefore, in the case of an undirected graph, we might instead compute the probability of drawing a path $i\rightsquigarrow j$ or $j\rightsquigarrow i$. The result is a symmetric matrix,

$$\mathbf{\Pi}_{\mathrm{sym}}=\mathbf{\Pi}+\mathbf{\Pi}^{\mathrm{T}} \tag{20}$$

and only the upper (or lower) triangular part of the matrix is relevant.

3.2.3. An intuitive interpretation of the z_{ij} in terms of killed random walk

An intuitive interpretation of the elements z_{ij} of the $\mathbf{Z}$ matrix can be provided as follows [87, 72]. Consider a special random walk defined by the transition probability matrix $\mathbf{W}$ whose elements are $[\mathbf{W}]_{ij} = p_{ij}^{\mathrm{ref}} \exp[-c_{ij}]$. As $\mathbf{W}$ has some row sums less than one (the rows i of $\mathbf{C}$ containing at least one strictly positive cost c_{ij}), the random walker has a nonzero probability of disappearing in each of these nodes which is equal to $(1 - \sum_{j=1}^{n} w_{ij})$ at each time step.

Indeed, from Equation (6), it can be observed that the probability of surviving during a transition $i \to j$ is proportional to $\exp[-\theta c_{ij}]$, which makes sense: there is a smaller probability to survive edges with a high cost. In this case, the elements of the $\mathbf{Z}$ matrix, $z_{ij} = [\mathbf{Z}]_{ij}$, can be interpreted as the expected number of times that a "killed" random walk, starting from node i, visits node j (see for instance [30, 56]) before being killed. This special stochastic process has been called an "evaporating random walk" in [87] or an "exponentially killed random walk" in [95].

4. Working with hitting/absorbing paths: the bag of hitting paths

The bag-of-hitting-paths model described in this section is a restriction of the previously introduced bag-of-paths model in which the ending node of each path only appears once – at the end of the path. In other words, no intermediate node on the path is allowed to be the ending node j, thus prohibiting looping on this node j. Technically this constraint will be enforced by making the ending node *absorbing*[6], as in the case of an absorbing Markov chain [30, 47, 56, 79]. We will see later in this section that this model has some nice properties.

4.1. Definition of the bag-of-hitting-paths probabilities

Let $\mathcal{P}_{ij}^{\mathrm{h}}$ be the set of *hitting* paths starting from i and stopping once node j has been reached for the first time (j is made absorbing and killing). Let $\mathcal{P}^{\mathrm{h}} = \cup_{ij}\mathcal{P}_{ij}^{\mathrm{h}}$ be the complete set of such hitting paths. Following the same reasoning as in the previous subsection, from Equation (9), when putting a Gibbs-Boltzmann distribution on $\mathcal{P}^{\mathrm{h}}$, the probability of drawing a hitting path starting in i and ending in j is

$$\mathrm{P_h}(s=i,e=j) = \frac{\sum\limits_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum\limits_{\wp'\in\mathcal{P}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp')\exp[-\theta\tilde{c}(\wp')]} = \frac{\sum\limits_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\mathcal{Z}_{\mathrm{h}}} \tag{21}$$

and the denominator of this expression is also called the **partition function**, $\mathcal{Z}_{\mathrm{h}} = \sum_{\wp\in\mathcal{P}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]$, for the hitting paths system this time. The quantity $\mathrm{P_h}(s=i,e=j)$ will be called the **bag-of-hitting-paths probability** of drawing a hitting path starting in i and ending in j. Note that in the case of unbounded hitting paths, the reference path probabilities can simply be defined as $\tilde{\mathrm{P}}^{\mathrm{ref}} = \frac{1}{n^2}\tilde{\pi}^{\mathrm{ref}}$ if we assume a uniform reference probability for drawing the

[6] And *killing*, see later.

starting and ending nodes. With this definition, it is shown in Appendix A that the probability is properly normalized, i.e., $\sum_{\wp\in\mathcal{P}^{\mathrm{h}}} \tilde{\mathrm{P}}^{\mathrm{ref}}(\wp) = 1$.

Obviously, for hitting paths, if we adopt the convention that zero-length paths are allowed, paths of length greater than 0 starting in node i and ending in the same node i are prohibited – in that case, the zero-length path is the only allowed path starting and ending in i and we set its $\tilde{\pi}^{\mathrm{ref}}$ equal to 1.

Now, following the same reasoning as in previous section, the numerator of Equation (21) is

$$\begin{aligned}\sum_{\wp\in\mathcal{P}^{\mathrm{h}}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)] &= \mathbf{e}_i^{\mathrm{T}}\left(\sum_{t=0}^{\infty}(\mathbf{W}^{(-j)})^t\right)\mathbf{e}_j = \mathbf{e}_i^{\mathrm{T}}(\mathbf{I}-\mathbf{W}^{(-j)})^{-1}\mathbf{e}_j \\ &= \mathbf{e}_i^{\mathrm{T}}\mathbf{Z}^{(-j)}\mathbf{e}_j = z_{ij}^{(-j)} \end{aligned} \tag{22}$$

where $\mathbf{W}^{(-j)}$ is now matrix $\mathbf{W}$ of Equation (6) where the jth row has been set to $\mathbf{0}^{\mathrm{T}}$ (node i is absorbing and killing meaning that the jth row of the transition matrix, $\mathbf{P}^{\mathrm{ref}}$, is equal to zero) and $\mathbf{Z}^{(-j)} = (\mathbf{I}-\mathbf{W}^{(-j)})^{-1}$. This means that when the random walker reaches node j, it immediately stops its walk there. In other words, the walker can only follow hitting paths, as required. This matrix is given by $\mathbf{W}^{(-j)} = \mathbf{W} - \mathbf{e}_j(\mathbf{w}_j^r)^{\mathrm{T}}$ with $\mathbf{w}_j^r = \mathbf{col}_j(\mathbf{W}^{\mathrm{T}}) = \mathbf{W}^{\mathrm{T}}\mathbf{e}_j$ being a column vector containing the jth row of $\mathbf{W}$.

4.2. Computation of the bag-of-hitting-paths probabilities

In Appendix B, it is shown from a bag-of-paths framework point of view that the elements of $\mathbf{Z}^{(-j)}$ can be computed simply and efficiently by

$$z_{ij}^{(-j)} = [\mathbf{Z}^{(-j)}]_{ij} = \frac{z_{ij}}{z_{jj}} \tag{23}$$

which is a noteworthy result by itself. Note that this result has been re-derived in a more conventional, but also more tedious, way through the Sherman-Morrison formula by [58] in the context of computing randomized shortest paths dissimilarities in closed form.

Using this result, Equation (22) can be developed as

$$\sum_{\wp\in\mathcal{P}^{\mathrm{h}}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)] = z_{ij}^{(-j)} = \frac{z_{ij}}{z_{jj}} \triangleq z_{ij}^{\mathrm{h}} \tag{24}$$

where we define the matrix containing the elements $z_{ij}^{(-j)} = z_{ij}/z_{jj}$ as $\mathbf{Z}_{\mathrm{h}}$ – the **fundamental matrix** for hitting paths. The elements of the matrix $\mathbf{Z}_{\mathrm{h}}$ are denoted by z_{ij}^{h}. From Equation (24), this matrix can be computed as $\mathbf{Z}_{\mathrm{h}} = \mathbf{Z}\mathbf{D}_{\mathrm{h}}^{-1}$ with $\mathbf{D}_{\mathrm{h}} = \mathbf{Diag}(\mathbf{Z})$. Note that the diagonal elements of $\mathbf{Z}_{\mathrm{h}}$ are equal to 1, $z_{ii}^{\mathrm{h}} = 1$. Moreover, when $\theta \to \infty$, $z_{jj} \to 1$ and $z_{ij}^{\mathrm{h}} \to z_{ij}$ (at the limit, only shortest paths, without loops, are considered).

From these results, we immediately deduce the bag-of-hitting-paths proba-

bility including zero-length paths (Equation (21)),

$$\mathrm{P_h}(s=i,e=j) = \frac{\displaystyle\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\displaystyle\sum_{i',j'=1}^{n}\sum_{\wp'\in\mathcal{P}_{i'j'}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp')\exp[-\theta\tilde{c}(\wp')]}$$
$$= \frac{z_{ij}/z_{jj}}{\displaystyle\sum_{i',j'=1}^{n}(z_{i'j'}/z_{j'j'})} = \frac{z_{ij}^{\mathrm{h}}}{\mathcal{Z}_{\mathrm{h}}} \tag{25}$$

where the denominator of Equation (25) is the partition function of the hitting paths model,

$$\mathcal{Z}_{\mathrm{h}} = \sum_{i,j=1}^{n}\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)] = \sum_{i,j=1}^{n}(z_{ij}/z_{jj}) \tag{26}$$

In matrix form, denoting by $\mathbf{\Pi}_{\mathrm{h}}$ the matrix of **bag-of-hitting-paths probabilities** $\mathrm{P_h}(s=i,e=j)$,

$$\mathbf{\Pi}_{\mathrm{h}} = \frac{\mathbf{Z}\mathbf{D}_{\mathrm{h}}^{-1}}{\mathbf{e}^{\mathrm{T}}\mathbf{Z}\mathbf{D}_{\mathrm{h}}^{-1}\mathbf{e}},\ \text{with}\ \mathbf{Z}=(\mathbf{I}-\mathbf{W})^{-1}\ \text{and}\ \mathbf{D}_{\mathrm{h}}=\mathbf{Diag}(\mathbf{Z}) \tag{27}$$

The algorithm for computing the matrix $\mathbf{\Pi}_{\mathrm{h}}$ is shown in Algorithm 1. The symmetric version for hitting paths is obtained by applying Equation (20) after the computation of $\mathbf{\Pi}_{\mathrm{h}}$. An interesting application would be to investigate graph cuts based on bag-of-hitting-paths probabilities instead of the standard adjacency matrix.

4.3. An intuitive interpretation of the z_{ij}^{h} in terms of killed random walk

In this subsection, we provide an intuitive description of the elements of the hitting paths fundamental matrix, $\mathbf{Z}_{\mathrm{h}}$. As for non-hitting paths, let us consider a special killed random walk with absorbing state α on the graph G whose transition probabilities are given by the elements of $\mathbf{W}^{(-j)}$, that is, $w_{ij} = p_{ij}^{\mathrm{ref}}\exp[-\theta c_{ij}]$ when $i\neq\alpha$ and $w_{\alpha j}=0$ otherwise. In other words, the node α is made *absorbing* and *killing* – it corresponds to hitting paths with node α as hitting node. When the walker reaches this node, he stops his walk and disappears. Moreover, as $\exp[-\theta c_{ij}]\leq 1$ for all i,j, the matrix of transition probabilities w_{ij} is substochastic and the random walker has also a nonzero probability $(1-\sum_{j=1}^{n}w_{ij})$ of disappearing at each step of its random walk and in each node i for which $(1-\sum_{j=1}^{n}w_{ij})>0$.

Now, let us consider column α (corresponding to the hitting, or absorbing, node) of the fundamental matrix of non-hitting paths, $\mathbf{col}_\alpha(\mathbf{Z}) = \mathbf{Z}\mathbf{e}_\alpha$. Because the fundamental matrix is $\mathbf{Z}=(\mathbf{I}-\mathbf{W})^{-1}$ (Equation (13)), we easily obtain $(\mathbf{I}-\mathbf{W})(\mathbf{Z}\mathbf{e}_\alpha)=\mathbf{I}\mathbf{e}_\alpha=\mathbf{e}_\alpha$. Or, in elementwise form,

$$\begin{cases} z_{i\alpha} = \sum_{j=1}^{n} w_{ij}z_{j\alpha} & \text{for each } i\neq\alpha \\ z_{\alpha\alpha} = \sum_{j=1}^{n} w_{\alpha j}z_{j\alpha}+1 & \text{for absorbing node } \alpha \end{cases} \tag{28}$$

Algorithm 1 Computing the bag-of-hitting-paths probability matrix of a graph.

Input:

– A weighted, possibly directed, strongly connected, graph G containing n nodes.
– The $n \times n$ adjacency matrix $\mathbf{A}$ associated to G, containing affinities.
– The $n \times n$ cost matrix $\mathbf{C}$ associated to G.
– The inverse temperature parameter θ.

Output:

– The $n \times n$ bag-of-hitting-paths probability matrix $\mathbf{\Pi}_{\mathrm{h}}$ containing the probability of drawing a path starting in node i and ending in node j, when sampling paths according to a Gibbs-Boltzmann distribution.

1. $\mathbf{D} \leftarrow \mathbf{Diag}(\mathbf{Ae})$ ▷ the outdegree diagonal matrix with $\mathbf{e}$ being a column vector full of 1's
2. $\mathbf{P}^{\mathrm{ref}} \leftarrow \mathbf{D}^{-1}\mathbf{A}$ ▷ the reference transition probabilities matrix
3. $\mathbf{W} \leftarrow \mathbf{P}^{\mathrm{ref}} \circ \exp[-\theta\mathbf{C}]$ ▷ elementwise exponential and multiplication $\circ$
4. $\mathbf{Z} \leftarrow (\mathbf{I} - \mathbf{W})^{-1}$ ▷ the fundamental matrix
5. $\mathbf{D}_{\mathrm{h}} \leftarrow \mathbf{Diag}(\mathbf{Z})$ ▷ the column-normalization diagonal matrix needed for computing hitting paths probabilities
6. $\mathbf{Z}_{\mathrm{h}} \leftarrow \mathbf{Z}\mathbf{D}_{\mathrm{h}}^{-1}$ ▷ column-normalize the fundamental matrix
7. $\mathcal{Z}_{\mathrm{h}} \leftarrow \mathbf{e}^{\mathrm{T}}\mathbf{Z}_{\mathrm{h}}\mathbf{e}$ ▷ compute the partition function
8. $\mathbf{\Pi}_{\mathrm{h}} \leftarrow \dfrac{\mathbf{Z}_{\mathrm{h}}}{\mathcal{Z}_{\mathrm{h}}}$ ▷ the bag-of-hitting-paths probability matrix
9. **return** $\mathbf{\Pi}_{\mathrm{h}}$

When considering hitting paths instead, $z^{\mathrm{h}}_{\alpha\alpha} = 1$ (see Equation (24)) because $w_{\alpha j} = 0$ for all j (node α is made absorbing and killing) so that the second line of Equation (28) – the boundary condition – becomes simply $z^{\mathrm{h}}_{\alpha\alpha} = 1$ for hitting paths. Moreover, we know that $z^{\mathrm{h}}_{i\alpha} = z_{i\alpha}/z_{\alpha\alpha}$ for any $i \neq \alpha$. Thus, dividing the first line of Equation (28) by $z_{\alpha\alpha}$ provides

$$\begin{cases} z^{\mathrm{h}}_{i\alpha} = \sum_{j=1}^{n} w_{ij}\, z^{\mathrm{h}}_{j\alpha} & \text{for each } i \neq \alpha \\ z^{\mathrm{h}}_{\alpha\alpha} = 1 & \text{for absorbing node } \alpha \end{cases} \tag{29}$$

Interestingly, this is exactly the set of recurrence equations computing the probability of hitting node α when starting from node i (see, e.g., [56, 86, 102]). Therefore, the $z^{\mathrm{h}}_{i\alpha}$ represent the *probability of surviving* during the killed random walk from i to α with transition probabilities w_{ij} and node α made absorbing. Said differently, it corresponds to the probability of reaching absorbing node j without being killed during the walk.

5. Two novel families of distances based on hitting paths probabilities

In this section, two families of distance measures are derived from the hitting paths probabilities including zero-length paths[7]. The second one benefits from some nice properties that will be detailed.

[7]Note that the results do not hold for a bag of paths excluding zero-length paths. Furthermore, the distances are developed only for the hitting paths case because the equivalent definitions applied to non-hitting paths are less satisfying or even ill-defined.

5.1. A first distance measure: the surprisal distance

The first distance measure is directly derived from the bag-of-paths probabilities introduced in the previous section.

5.1.1. Definition of the distance

This section shows that the associated (directed) surprisal measure,

$$-\log \mathrm{P_h}(s = i, e = j),$$

quantifying the "surprise" generated by the outcome $(s = i) \wedge (e = j)$, when symmetrized, is a distance measure. This distance $\Delta_{ij}^{\mathrm{sur}}$ associated to the bag-of-hitting-paths is defined as follows

$$\Delta_{ij}^{\mathrm{sur}} \triangleq \begin{cases} -\dfrac{\log \mathrm{P_h}(s = i, e = j) + \log \mathrm{P_h}(s = j, e = i)}{2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{30}$$

where $\mathrm{P_h}(s = i, e = j)$ and $\mathrm{P_h}(s = j, e = i)$ are computed according to Equation (25) or (27) for the matix form. Obviously, $\Delta_{ij}^{\mathrm{sur}} \geq 0$ and $\Delta_{ij}^{\mathrm{sur}}$ is symmetric. Moreover, $\Delta_{ij}^{\mathrm{sur}}$ is equal to zero if and only if $i = j$.

It is shown in Appendix C that this quantity is a distance measure since it satisfies the triangle inequality, in addition to the other mentioned properties. This distance will be called the bag-of-hitting-paths **surprisal distance**.

5.1.2. Computation of the distance

The surprisal distance can easily be computed by adding the following matrix operations to Algorithm 1:

- $\mathbf{\Delta}_{\mathrm{sur}} \leftarrow -\frac{1}{2}\Big[\log(\mathbf{\Pi}_{\mathrm{h}}) + \log(\mathbf{\Pi}_{\mathrm{h}}^{\mathrm{T}})\Big]$ ▷ take elementwise logarithm for computing the surprisal distances
- $\mathbf{\Delta}_{\mathrm{sur}} \leftarrow \mathbf{\Delta}_{\mathrm{sur}} - \mathbf{Diag}(\mathbf{\Delta}_{\mathrm{sur}})$ ▷ put diagonal to zero

We now turn to the development of the second distance measure.

5.2. A second distance measure: the potential distance

This subsection introduces a second measure enjoying some nice properties, based on the same ideas. Note that the same distance also appeared in [58], where a distance closely related to the randomized shortest paths dissimilarity [87, 113] was derived from a different perspective (minimizing free energy), and called the free energy distance in that work.

5.2.1. Definition of the distance

The second distance measure automatically follows from Inequality (C.5) in Appendix C and is based on the quantity $\phi(i, j) = -\frac{1}{\theta} \log z_{ij}^{\mathrm{h}}$, which is necessarily nonnegative because z_{ij}^{h} can be interpreted as a probability (see Subsection 4.3). For convenience, let us recall this inequality,

$$\mathrm{P_h}(s = i, e = k) \geq \mathcal{Z}_{\mathrm{h}} \mathrm{P_h}(s = i, e = j)\, \mathrm{P_h}(s = j, e = k)$$

Then, from $\mathrm{P_h}(s=i,e=j) = z^{\mathrm{h}}_{ij}/\mathcal{Z}_{\mathrm{h}}$ (Equation (25)), we directly obtain $z^{\mathrm{h}}_{ik} \geq z^{\mathrm{h}}_{ij}\, z^{\mathrm{h}}_{jk}$. Taking $-\frac{1}{\theta}\log$ of both sides provides $-\frac{1}{\theta}\log z^{\mathrm{h}}_{ik} \leq -\frac{1}{\theta}\log z^{\mathrm{h}}_{ij} - \frac{1}{\theta}\log z^{\mathrm{h}}_{jk}$, or,

$$\phi(i,k) \leq \phi(i,j) + \phi(j,k) \tag{31}$$

where we defined

$$\phi(i,j) \triangleq -\frac{1}{\theta}\log z^{\mathrm{h}}_{ij} = -\frac{1}{\theta}\log\left(\frac{z_{ij}}{z_{jj}}\right) \tag{32}$$

and, from (31), the $\phi(i,j)$ verify the triangle inequality.

The quantity $\phi(i,j)$ will be called the directed *potential* [23] of node i with respect to node j. Indeed, it has been shown [38] that when computing the continuous-state continuous-time equivalent of the randomized shortest paths framework [87], $\phi(x,y)$ plays the role of a potential inducing a drift (external force) $\boldsymbol{\nabla}\phi$ in the corresponding diffusion equation.

From the properties and the probabilistic interpretation of the z^{h}_{ij}, both $\phi(i,j) \geq 0$ (as $0 \leq z^{\mathrm{h}}_{ij} \leq 1$) and $\phi(i,i) = 0$ (as $z^{\mathrm{h}}_{ii} = 1$) hold. This directed distance measure has three intuitive interpretations.

- First, let us recall from Equation (24) that z^{h}_{ij} is given by $z^{\mathrm{h}}_{ij} = \sum_{\wp\in\mathcal{P}^{\mathrm{h}}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)] = z_{ij}/z_{jj}$ where z_{ij} is element i,j of the fundamental matrix $\mathbf{Z}$ (see Equation (13)). From this last expression, $\phi(i,j)$ can be interpreted (up to a scaling factor) as the logarithm of the expectation of the reward $\exp[-\theta\tilde{c}(\wp)]$ with respect to the path likelihoods, when considering absorbing random walks starting from node i and ending in node j.

- In addition, from Equation (29), it also corresponds to minus the log-likelihood of surviving during the killed, absorbing, random walk from i to j.

- Finally, it was shown in [58], investigating further developments of the randomized shortest paths (RSP) dissimilarity, that the potential distance also corresponds to the minimal free energy of the system of hitting paths from i to j. Indeed, the RSP dissimilarity, defined as the expected total cost between i and j, is not a distance measure as it does not satisfy the triangle inequality. However, subtracting the entropy multiplied by the temperature from the expected total cost (that is, computing the free energy) leads to a distance measure that was shown to be equivalent to the potential distance. Therefore the potential distance was called the **free energy distance** in [58], which provides still another interpretation to the potential distance.

Inequality (31) suggests to define the distance $\Delta^{\phi}_{ij} = (\phi(i,j) + \phi(j,i))/2$. It has all the properties of a distance measure, including the triangle inequality, which is verified thanks to Inequality (31). Note that this distance measure can be expressed as a function of the surprisal distance (see Equation (30)) as $\Delta^{\phi}_{ij} = (\Delta^{\mathrm{sur}}_{ij} - \log\mathcal{Z}_{\mathrm{h}})/\theta$ for $i \neq j$. This shows that the newly introduced distance is equivalent to the previous one, up to the addition of a constant and a rescaling.

The definition of the bag-of-hitting-paths (free energy) **potential distance** is therefore

$$\Delta_{ij}^{\phi} \triangleq \begin{cases} \dfrac{\phi(i,j)+\phi(j,i)}{2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}, \text{ where } \phi(i,j) = -\frac{1}{\theta}\log\left(\frac{z_{ij}}{z_{jj}}\right) \tag{33}$$

and z_{ij} is element i,j of the fundamental matrix $\mathbf{Z}$ (see Equation (13)).

5.2.2. Computation of the distance

From Equation (27), it can be easily seen that the matrix $\mathbf{Z}_{\mathrm{h}}$ containing the z_{ij}^{h} can be computed thanks to Algorithm 1 without the normalization steps 7 and 8. The distance matrix with elements Δ_{ij}^{ϕ} is denoted as $\mathbf{\Delta}_{\phi}$ and can easily be obtained by adding the following matrix operations to Algorithm 1:

- $\mathbf{\Phi} \leftarrow -\log(\mathbf{Z}_{\mathrm{h}})/\theta$ ▷ take elementwise logarithm for computing the potentials
- $\mathbf{\Delta}_{\phi} \leftarrow (\mathbf{\Phi} + \mathbf{\Phi}^{\mathrm{T}})/2$ ▷ symmetrize the matrix
- $\mathbf{\Delta}_{\phi} \leftarrow \mathbf{\Delta}_{\phi} - \mathbf{Diag}(\mathbf{\Delta}_{\phi})$ ▷ put diagonal to zero

Note that both the surprisal and the potential distances are well-defined as we assumed that G is strongly connected.

5.3. Some properties of the potential and surprisal distances

The potential distance Δ^{ϕ} benefits from some interesting properties proved in the appendix:

- **The potential distance is cutpoint additive**, meaning that $\Delta_{ik}^{\phi} = \Delta_{ij}^{\phi} + \Delta_{jk}^{\phi}$ if and only if every path from i to k passes through node j [15] (see Appendix D for the proof).
- **For an undirected graph G, the distance Δ_{ij}^{ϕ} approaches the shortest-path distance** when θ becomes large, $\theta \to \infty$. In that case, the Equation (33) reduces to the Bellman-Ford formula (see, e.g., [7, 20, 25]) for computing the shortest-path distance, $\Delta_{ik}^{\mathrm{SP}} = \min_{j \in \mathcal{S}ucc(i)}\{c_{ij} + \Delta_{jk}^{\mathrm{SP}}\}$ and $\Delta_{kk}^{\mathrm{SP}} = 0$ (see Appendix E for the proof). The convergence is, however, slow[8] and numerical underflows could appear before complete convergence to the shortest-path distances (convergence is linear in θ – see the appendix for details). Therefore, if solutions close to the shortest-path distance are needed (with very large θ), computational tricks such as those used in hidden Markov models should be implemented. See for instance the appendix in [46].
- **For an undirected graph G, the distance Δ_{ij}^{ϕ} approaches half the commute-cost distance** when θ becomes small, $\theta \to 0^{+}$ (see Appendix F for the proof). Note that, for a given graph G, the commute cost between two nodes is proportional to the commute time between these two nodes, and therefore also proportional to the resistance distance (see [12, 58]).

[8] It was observed, e.g., that the convergence of the RSP dissimilarity [87, 113] is much faster when θ increases.

- **The distance Δ_{ij}^{ϕ} extends the Bellman-Ford formula computing the shortest-path distance to integrate sub-optimal paths (exploration)** by simply replacing the min operator by the softmin operator in the recurrence formula. This property is discussed in the next subsection.

All of these properties make the potential distance quite attractive as it defines a family of distances interpolating between the shortest-path and the resistance distance. Our conjecture is that interpolating between these two distances hopefully alleviates the "lost in space" effect [106, 107] as the distance gradually focuses on shorter paths, while still exploring sub-optimal paths, when parameter θ increases. A recent paper [43] addresses this question by showing the consistency and the robustness of the Laplacian transformed hitting time (the Laplace transform of hitting times), a measure related to the potential distance. One of our future work will be to evaluate if their analysis can be transposed to our measures. But, of course, ultimately, the "best" distance is application- and data-dependent and it is difficult to know in advance which one will perform best.

Note that, even if the potential distance converges to the commute cost when $\theta \to 0^+$, we have to stress that θ should not become equal to zero because the matrix $\mathbf{W}$ becomes rank-deficient when $\theta = 0$. This means that the Equation (13) cannot be used for computing the commute cost when θ is *exactly* equal to zero. Despite this annoying fact, we found that the approximation is quite accurate for small values of θ.

Concerning the surprisal distance, because it was shown in the previous section that $\Delta_{ij}^{\text{sur}} = \theta \Delta_{ij}^{\phi} + \log \mathcal{Z}_{\text{h}}$ for all $i \neq j$, we deduce that the ranking of the node distances for a given θ is the same for the two distances.

5.4. Relationships with the Bellman-Ford formula

As shown in Appendix E (Equation (E.7)) the potential $\phi(i,k)$ for a fixed ending node k can be computed thanks to the following recurrence formula

$$\phi(i,k) = \begin{cases} -\dfrac{1}{\theta} \log \left[\displaystyle\sum_{j \in \mathcal{S}ucc(i)} p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \phi(j,k))] \right] & \text{if } i \neq k \\ 0 & \text{if } i = k \end{cases} \tag{34}$$

which is an extension of Bellman-Ford's formula for computing the shortest-path distance in a graph [7, 20, 25, 52, 85, 92]. The Equation (34) has to be iterated until convergence. Note that this result seems to be related to the concept of "path integral control" developed in control theory; see, e.g., the survey [54]. The nonlinear recurrence (34) is also a generalization of the distributed consensus algorithm developed in [98], and considering binary costs only.

Interestingly and intriguingly, this expression is obtained by simply replacing the min operator by a weighted version of the softmin operator [24] in the Bellman-Ford recurrence formula,

$$\text{softmin}_{\mathbf{q},\theta}(\mathbf{x}) = -\frac{1}{\theta} \log \left(\sum_{j=1}^{n} q_j \exp[-\theta x_j] \right) \text{ with all } q_j \geq 0 \text{ and } \textstyle\sum_{j=1}^{n} q_j = 1 \tag{35}$$

which interpolates between weighted average and minimum operators (see Appendix E or [24, 98]). The consequence is that the potential $\phi(i,j)$ tends to the average first-passage cost when $\theta \to 0^+$ and to the shortest-path cost when $\theta \to \infty$ (see Appendix E).

6. Extending the bag of paths by considering non-uniform priors on nodes

This section extends the bag of hitting paths model by considering non-uniform a priori probabilities of selecting the starting and ending nodes[9]. For instance, if the nodes represent cities, it could be natural to weigh each city by its population. These prior probabilities, weighting each node of G, will be denoted as q_i^{s} and q_j^{e} with $\sum_{i=1}^n q_i^{\mathrm{s}} = 1$, $\sum_{j=1}^n q_j^{\mathrm{e}} = 1$ and all weights non-negative.

In this situation, because the reference probability $\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp_{ij})$ becomes

$$\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp_{ij}) = q_i^{\mathrm{s}} q_j^{\mathrm{e}}\, \tilde{\pi}^{\mathrm{ref}}(\wp_{ij}), \tag{36}$$

instead of $\frac{1}{n^2}\tilde{\pi}^{\mathrm{ref}}(\wp_{ij})$, the probability of sampling a hitting path $i \rightsquigarrow j$ in Equation (21) is redefined as

$$\begin{aligned}
\mathrm{P}_{\mathrm{h}}(s=i,e=j) =& \frac{\displaystyle\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\mathrm{P}}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\displaystyle\sum_{\wp'\in\mathcal{P}^{\mathrm{h}}} \tilde{\mathrm{P}}^{\mathrm{ref}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \\
=& \frac{q_i^{\mathrm{s}}\left(\displaystyle\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]\right) q_j^{\mathrm{e}}}{\displaystyle\sum_{i',j'=1}^{n} q_{i'}^{\mathrm{s}}\left(\sum_{\wp'\in\mathcal{P}_{i'j'}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp')\exp[-\theta\tilde{c}(\wp')]\right) q_{j'}^{\mathrm{e}}}
\end{aligned} \tag{37}$$

where $\tilde{\pi}^{\mathrm{ref}}(\wp_{ij})$ is, as before, the likelihood of the path $\wp_{ij}$ given that the starting and ending nodes are i, j. Therefore this expression can be computed thanks to Equation (24) as the weighted quantity

$$\mathrm{P}_{\mathrm{h}}(s=i,e=j) = \frac{q_i^{\mathrm{s}}\left(\dfrac{z_{ij}}{z_{jj}}\right) q_j^{\mathrm{e}}}{\displaystyle\sum_{i',j'=1}^{n} q_{i'}^{\mathrm{s}}\left(\frac{z_{i'j'}}{z_{j'j'}}\right) q_{j'}^{\mathrm{e}}} = \frac{q_i^{\mathrm{s}}\left(\dfrac{z_{ij}}{z_{jj}}\right) q_j^{\mathrm{e}}}{\mathcal{Z}_{\mathrm{hw}}} \tag{38}$$

and the denominator

$$\mathcal{Z}_{\mathrm{hw}} = \sum_{i,j=1}^{n} q_i^{\mathrm{s}}\left(\frac{z_{ij}}{z_{jj}}\right) q_j^{\mathrm{e}} \tag{39}$$

[9] The development for non-hitting paths is similar and will therefore be omitted.

is the new, weighted by priors, partition function. The numerator of (38) defines the fundamental matrix of the hitting paths system for weighted nodes, containing elements

$$q_i^{\mathrm{s}} \left(\frac{z_{ij}}{z_{jj}} \right) q_j^{\mathrm{e}} = q_i^{\mathrm{s}} \, z_{ij}^{\mathrm{h}} \, q_j^{\mathrm{e}} \quad \text{where, as before, } z_{ij}^{\mathrm{h}} = \frac{z_{ij}}{z_{jj}} \tag{40}$$

In matrix form, the counterpart of Equation (27) – but now including priors on the nodes – is

$$\mathbf{\Pi}_{\mathrm{h}} = \frac{\mathbf{Diag}(\mathbf{q}_{\mathrm{s}})\mathbf{Z}\mathbf{D}_{\mathrm{h}}^{-1}\mathbf{Diag}(\mathbf{q}_{\mathrm{e}})}{\mathbf{q}_{\mathrm{s}}^{\mathrm{T}}\mathbf{Z}\mathbf{D}_{\mathrm{h}}^{-1}\mathbf{q}_{\mathrm{e}}}, \text{ with } \mathbf{D}_{\mathrm{h}} = \mathbf{Diag}(\mathbf{Z}) \tag{41}$$

where the vectors $\mathbf{q}_{\mathrm{s}}$ and $\mathbf{q}_{\mathrm{e}}$ contain the a priori probabilities q_i^{s} and q_i^{e}. Of course, we recover Equation (27) when $\mathbf{q}_{\mathrm{s}} = \mathbf{q}_{\mathrm{e}} = \mathbf{e}/n$.

Interestingly, the surprisal and potential distances defined on the weighted nodes still verify the triangle inequality and are therefore distance measures; this is shown in Appendix G. Therefore, both the surprisal and the potential distances are defined in the same way as in previous section (see Equations (30) and (33)), but based this time on the weighted quantities defined in Equations (38) and (40).

More precisely, the directed surprisal distance is computed by taking $-\log$ of the probabilities (38) or (41) (matrix form) while the directed potential distance is redefined as (see Appendix G for details)

$$\phi(i,j) = -\frac{1}{\theta} \log(q_i^{\mathrm{s}} z_{ij}^{\mathrm{h}} q_j^{\mathrm{e}}) \tag{42}$$

7. Experiments on semi-supervised classification tasks

This experimental section aims at investigating the performance of the bag-of-hitting-paths distances, and kernels derived from them, in a semi-supervised classification task, on which they are compared with other competitive techniques.

Notice, however, that the goal of this experiment is not to design a state-of-the-art classifier. Rather, the main objective is to study the results of the proposed distances in comparison with other measures and therefore investigate their usefulness in solving pattern recognition tasks. More precisely, this experiment investigates to which extent the distance measures are able to accurately capture the global structure of the graph through a spectral method.

7.1. Graph based semi-supervised classification

Semi-supervised graph node classification has received an increasing interest in recent years (see [1, 13, 45, 96, 121, 122] for surveys). It considers the task of using the graph structure and other available information for inferring the class labels of unlabeled nodes of a network in which only a part of the class labels of nodes are known a priori. Several categories of approaches have been suggested for this problem. Among them, we may mention random walks [120, 97, 11], graph mincuts [8], spectral methods [14, 94, 62, 53], regularization frameworks [5, 108, 110, 118, 119], transductive and spectral support vector machines (SVM) [51], to name a few.

Topic	Size	Topic	Size	Topic	Size
news-2cl-1		**news-2cl-2**		**news-2cl-3**	
Politics/general	200	Computer/graphics	200	Space/general	200
Sport/baseball	200	Motor/motorcycles	200	Politics/mideast	200
news-3cl-1		**news-3cl-2**		**news-3cl-3**	
Sport/baseball	200	Computer/windows	200	Sport/hockey	200
Space/general	200	Motor/autos	200	Religion/atheism	200
Politics/mideast	200	Religion/general	200	Medicine/general	200
news-5cl-1		**news-5cl-2**		**news-5cl-3**	
Computer/windowsx	200	Computer/graphics	200	Computer/machardware	200
Cryptography/general	200	Computer/pchardware	200	Sport/hockey	200
Politics/mideast	200	Motor/autos	200	Medicine/general	200
Politics/guns	200	Religion/atheism	200	Religion/general	200
Religion/christian	200	Politics/mideast	200	Forsale/general	200

Table 2: Document subsets for semi-supervised classification experiments. Nine subsets have been extracted from the original 20 Newsgroups dataset, with 2, 3 and 5 topics as proposed in [111, 112]. Each class is composed of 200 documents.

Still another family of approaches is based on kernel methods, which embed the nodes of the input graph into a Euclidean feature space where a decision boundary can be estimated using standard kernel (semi-)supervised methods, such as SVMs. Fouss et al. [34] investigated the applicability of nine such graph kernels in collaborative recommendation and semi-supervised classification by adopting a simple sum-of-similarities[10] rule (SoS). Zhang et al. [117, 116] as well as Tang et al. [99, 100, 101] extract the dominant eigenvectors (a "latent space") of graph kernels or similarity matrices and then input them to a supervised classification method, such as a logistic regression or a SVM, to categorize the nodes. These techniques based on similarities and eigenvectors extraction allow to scale to large graphs, depending on the kernel.

Another category of classification methods relies on random walks performed on a weighted and possibly directed graph seen as a Markov chain. The random walk with restart [80, 104, 105], directly inspired by the PageRank algorithm, is one of them. The method of Callut et al. [11], based on discriminative random walks, or $\mathcal{D}$-walks, belongs to the same category. It defines, for each class, a group betweenness measure based on passage times during special random walks of bounded length. Those walks are constrained to start and end in nodes within the same class, defining distinct random walks for each class. The number of passages on nodes is computed for each type of such random walk, therefore defining a distinct betweenness for each class. The main advantage of some of these random walk-based approaches is that class labels can be computed efficiently (in linear time) while providing competitive results.

7.2. Datasets description

Comparison of the different methods will be performed on several well-known real world graph datasets (14 in total). Note that, in some cases, only the largest connected component of the following graphs has been selected:

[10]The equivalent of nearest neighbors classification when dealing with similarities (a kernel matrix) instead of distances.

- **20 Newsgroups (9 subsets)**: this dataset[11] is composed of 20000 text documents taken from 20 discussion groups of the Usenet diffusion list (available on UCI [67]). Nine subsets related to different topics are extracted from the original dataset, as listed in Table 2 [111, 112]. Each subset is composed of 200 documents extracted randomly from the different newsgroups. The subsets with two classes (news-2cl-1,2,3) contain 400 documents, 200 in each class. In the same way, subsets with three classes contain 600 documents and subsets with five classes contain 1000 documents. Each subset is composed of different topics, each of which are either easy to separate (Computer/windowsx and Religion/christian) or harder to separate (Computer/graphics and Computer/pchardware). Initially, this dataset does not have a graph structure but is represented in a word vector space of high dimensionality. To transform this dataset into a graph structure, a fairly standard preprocessing has been performed, which is directly inspired by the paper of Yen et al. [112].

  Basically, the first step is to reduce the high dimensionality of the feature space (terms), by removing stop words, applying a stemming algorithm on each term, removing too common or uncommon terms and by removing terms with low mutual information with documents. Second, a term-document matrix $\mathbf{W}$ is constructed with the remaining terms and documents. The elements w_{ij} are *tf-idf* values [70] of term i in document j. Each row of the term-document matrix $\mathbf{W}$ is then normalized to 1. Finally, the adjacency matrix defining the links between documents is given by $\mathbf{A} = \mathbf{W}^{\mathbf{T}}\mathbf{W}$.

- **IMDB**: the collaborative Internet Movie Database (`IMDb`, [69]) has several applications such as making movie recommendations, clustering or movie category classification. It contains a graph of movies linked together whenever they share the same production company. The weight of an edge in the resulting graph is the number of production companies two movies have in common. The classification problem focuses on identifying clusters of movies that share the same notoriety (whether the movie is a box-office hit or not).

- **WebKB (4 datasets)**: these networks consist of sets of web pages gathered from four computer science departments (one for each university, [69]), with each page manually labeled into 6 categories: course, department, faculty, project, staff, and student. Two pages are linked by co-citation (if x links to z and y links to z, then x and y are co-citing z).

The adjacency matrices extracted from these datasets are all symmetric and some are weighted. In a standard way, the costs associated to the edges are set to $c_{ij} = 1/a_{ij}$. That is, the elements of the adjacency matrix are considered as conductances and the costs as resistances. For unweighted graphs, affinities and costs are both equal to 1 for existing edges, meaning that the paths are weighted by their total length (number of steps).

[11] Available, e.g., from http://people.csail.mit.edu/jrennie/20Newsgroups/.

7.3. Compared distances, kernels, and algorithms

This paper derived distance measures from the bag-of-paths probabilities. In order to use these distances in machine learning and pattern recognition methods, it is convenient to transform them into similarity matrices, simply called kernels for convenience.

7.3.1. Deriving a kernel from a distance

From classical multidimensional scaling (MDS, see, e.g., [9, 26]), a centered kernel matrix $\mathbf{K}$ can be derived from a matrix of squared distances $\boldsymbol{\Delta}^{(2)}$ as follows

$$\mathbf{K}^{\text{mds}} = -\frac{1}{2}\mathbf{H}\boldsymbol{\Delta}^{(2)}\mathbf{H} \tag{43}$$

where $\mathbf{H} = (\mathbf{I} - \mathbf{e}\mathbf{e}^T/n)$ is the centering matrix and matrix $\boldsymbol{\Delta}^{(2)}$ contains the elementwise squared distances. Then, computing the dominant eigenvectors of this matrix (see the next section on Experimental settings) corresponds exactly to classical multidimensional scaling.

Still another popular way to map the distance matrix to a kernel matrix aims to use the Gaussian mapping or kernel (see, e.g., [90])

$$\mathbf{K}^{\text{g}} = \exp\Big[-\boldsymbol{\Delta}^{(2)}/2\sigma^2\Big] \tag{44}$$

where the exponential is taken elementwise. Both approaches will be investigated. Computing the dominant eigenvectors of this matrix corresponds to a kernel principal components analysis [90, 91].

However, the obtained kernels are not necessarily positive semi-definite until the distance is Euclidean, which is required for kernel methods. This problem can be fixed by removing the negative eigenvalues in the spectral decomposition (see, e.g., [73]), which will be applied in all our experiments[12].

For classifying the nodes, the five dominant eigenvectors of the resulting kernels will be extracted and then injected into a SVM classifier (see the next Subsection 7.4 for details).

7.3.2. Compared methods

The following list presents the methods based on kernels computed from the distances introduced in this paper, as well as from two other recent families of dissimilarities, for comparison. The derived kernels are computed by using both (1) multidimensional scaling (**mds**, Equation (43)) and (2) a Gaussian kernel (**g**, Equation (44)).

- The kernels associated to the bag-of-hitting-paths potential distance ($\mathbf{K}^{\text{mds}}_{\text{BoPP}}$, $\mathbf{K}^{\text{g}}_{\text{BoPP}}$) (Equations (33) and (43)-(44)). The corresponding methods are denoted as **BoPP-mds** and **BoPP-g**.

- The kernels associated to the bag-of-hitting-paths surprisal distance ($\mathbf{K}^{\text{mds}}_{\text{BoPS}}$, $\mathbf{K}^{\text{g}}_{\text{BoPS}}$) (Equations (30)) and (43)-(44)). The corresponding methods are denoted as **BoPS-mds** and **BoPS-g**.

[12]Note that, probably because only the dominant eigenvectors are extracted, we did not observe any significant difference in the experimental results when removing and not removing the negative eigenvalues of the kernels (results not reported).

- The randomized shortest-path (RSP) kernel ($\mathbf{K}_{\mathrm{RSP}}^{\mathrm{mds}}$, $\mathbf{K}_{\mathrm{RSP}}^{\mathrm{g}}$) computed from the RSP family of dissimilarities (see [58, 113, 87] and Equations (43)-(44)). The corresponding methods are denoted as **RSP-mds** and **RSP-g**.

- The logarithmic forest (LF) kernel ($\mathbf{K}_{\mathrm{LF}}^{\mathrm{mds}}$, $\mathbf{K}_{\mathrm{LF}}^{\mathrm{g}}$) computed from the logarithmic forest family of distance (see [15, 16] and Equations (43)-(44)). The corresponding methods are denoted as **LF-mds** and **LF-g**.

In addition, five state-of-the-art similarity matrices and kernels on a graph are added to this list and compared to the previous ones. We selected the three kernels providing consistently the best results in [34], which were based on a sum-of-similarities instead of the spectral method investigated in this paper.

- The modularity matrix ($\mathbf{Q}$) [77, 78], which was used as a kernel for semi-supervised learning earlier by Zhang et al. [117, 116] as well as Tang et al. [99, 100, 101]. The modularity matrix performed best in their experiments, in comparison with other state-of-the-art methods. This is our *first baseline* method, denoted as **Q**.

- The Markov diffusion kernel ($\mathbf{K}_{\mathrm{MD}}$) [34] computed from the Markov diffusion map distance [75, 76, 81, 82] and studied in [114, 34]. This kernel, as well as the two following ones, provided good results in [34]. The corresponding method is denoted as **MD**.

- The regularized Laplacian, or matrix forest, kernel ($\mathbf{K}_{\mathrm{RL}}$) [48, 18, 19, 34]. The corresponding method is denoted as **RL**.

- The regularized commute-time kernel ($\mathbf{K}_{\mathrm{RCT}}$) [34, 71]. The corresponding method is denoted as **RCT**.

- The bag-of-paths modularity matrix ($\mathbf{K}_{\mathrm{BoPM}}$) studied in [28]. The corresponding method is denoted as **BoPM**.

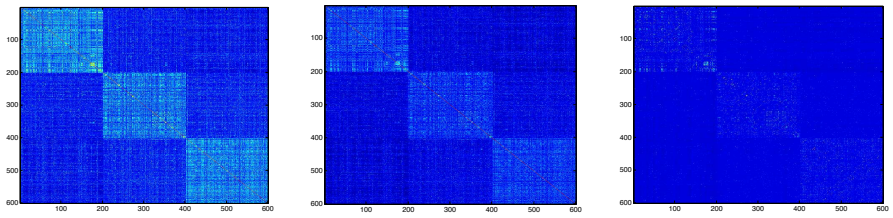
Finally, our introduced distances are also compared to an efficient, alternative, way of performing semi-supervised classification on a network:

- A sum-of-similarities (SoS) algorithm based on the regularized commute-time kernel, which provided good results on large datasets in [71]; see this paper for details. This is our *second baseline* method, denoted as **SoS**.

These kernels and similarity matrices are real symmetric when working with undirected graphs. All the above kernels and methods will be compared by following the experimental settings described hereafter. For illustration, a picture of some of the kernels is shown in Figure 1.

7.4. Experimental settings

In this experiment, we address the task of classification of unlabeled nodes in partially labelled graphs. The method we use is directly inspired from [99]. It consists of two steps:



(a) (b) (c)

Figure 1: Images of the different similarity matrices, (a) $\mathbf{K}^{\text{mds}}_{\text{BoPP}}$, (b) $\mathbf{K}^{\text{mds}}_{\text{BoPS}}$, and (c) $\mathbf{Q}$, computed on the news-3cl-1 dataset. Nodes have been sorted according to classes. We observe that classes are clearly visible in (a) and (b). For the standard modularity (c), the class discrimination is less clear.

1. Extracting the "latent social" dimensions, which may be done using any matrix decomposition technique or by using a graphical topic model. Here, we used, as in [99], a simple spectral decomposition of the relevant matrices. More precisely, we extracted the top eigenvectors of the compared kernel matrices just described (see Subection 7.3). This aims to perform a classical multidimensional scaling from distances when using the MDS transformation of Equation (43) and a kernel principal components analysis when using the Gaussian mapping of Equation (44).
2. Training a classifier on the extracted latent space. In this space, each feature corresponds to one latent variable (i.e. one of the top eigenvectors). The number of social dimensions has been set to 5 for all the suggested measures and the classifier is a one-vs-rest linear SVM. Note that we also investigated different numbers of social dimensions $[10, 50, 500]$ but the performance did not change significantly – these results are therefore not reported here.

The classification accuracy is computed for a labeling rate of 20%, i.e. proportion of nodes for which the label is known[13]. The labels of remaining nodes (80%) are removed and used as test data. For this considered labeling rate, an external stratified 5-fold cross-validation (each fold defining in turn the 20% labeled data) was performed, on which classification accuracies are averaged. For each fold of the external cross-validation, a 5-fold internal cross-validation is performed on the remaining labelled nodes in order to tune the hyper-parameters of the SVM and each kernel/distance ($\theta = \{0.01, 0.1, 1, 2, 5, 10\}$ for the bag-of-paths based approaches and $c = \{0.01, 0.1, 1, 10, 100\}$ for the SVM). Then, the performance obtained for each external fold is assessed on the remaining, unlabeled, nodes (test data) with the hyper-parameter tuned during the internal cross-validation.

For each unlabeled node, the various classifiers predict the most suitable category according to the procedure described below. We compute, for each method, the average classification accuracy obtained on the five folds of the

[13]Other settings were also investigated, leading to similar conclusions; they are therefore omitted here.

cross-validation. Then, a Borda ranking as well as a nonparametric Friedman-Nemenyi statistical test [27] are performed across all datasets in order to compare the different methods. Finally, pairwise comparisons between some specific methods are investigated.

7.5. Results and discussion

Table 3 reports average classification accuracies (in percent) of the methods on all the datasets, for a proportion of 20% of labeling rate and five latent dimensions (components). The method performing best is presented in boldface for each data set. Then, a simple Borda ranking of the methods is performed and shown in Table 4. Each method is given a score equal to its rank (methods are sorted in ascending order of accuracy, worst first and best last) for each dataset. The best method overall is the one showing the highest Borda score.

From these tables, it can be observed that the bag-of-paths (BoP) and the randomized shortest paths (RSP) based approaches obtain competitive results in comparison with the other methods. Indeed, both the BoPP and the BoPS consistently provide good results. The logarithmic forest distance also obtains good overall results. However, we can further observe that the best method is dataset-dependent; this shows that it is often useful to investigate different methods when facing a network-based semi-supervised classification problem. Moreover, the differences in performance among the best performing methods are often small. For instance, for the five best techniques (BoPP-g, BoPS-mds, BoPP-mds, RSP-g, BoPS-g; see Table 4), the average difference between the accuracy (see Table 3) of the best performing method and the other methods across all datasets is 0.79 and the maximum difference is only 4.58. This can be understood by the fact that we selected the most promising candidate methods for the comparisons, but also by the fact that several investigated distances are derived from a similar framework.

Method: Dataset:	BoPM	BoPP-g	BoPP-mds	BoPS-g	BoPS-mds	LF-g	LF-mds	MD	Q	RCT	RL	RSP-g	RSP-mds	SoS
webKB-texas	74.85	**77.40**	74.92	76.57	76.95	74.92	72.75	58.01	72.75	70.89	48.73	74.92	75.75	74.63
webKB-washington	66.19	71.49	70.68	68.61	70.05	**72.24**	70.10	66.53	59.50	67.40	40.78	70.68	70.33	65.61
webKB-wisconsin	72.84	**75.50**	73.49	73.13	74.14	73.78	71.91	70.48	72.70	70.76	45.04	73.35	72.49	73.71
webKB-cornell	**60.04**	55.57	58.31	56.29	58.46	58.38	56.43	51.73	51.23	46.82	41.91	58.31	56.87	58.67
imdb	74.44	50.75	50.68	50.77	50.68	50.71	50.71	52.67	66.64	56.93	68.44	50.75	50.68	**78.14**
news-2cl-1	96.00	95.06	94.25	95.25	94.75	95.06	95.94	**97.56**	94.81	90.94	90.69	94.31	94.06	92.50
news-2cl-2	89.83	91.02	90.70	**91.71**	91.58	90.89	89.26	90.64	91.02	86.43	87.50	90.52	90.89	89.89
news-2cl-3	94.49	**95.99**	95.68	95.80	95.99	95.55	95.05	95.49	94.17	92.86	93.36	95.99	95.30	94.11
news-3cl-1	**94.42**	93.92	93.08	92.92	93.08	93.17	92.25	91.75	93.33	72.17	78.25	93.50	92.67	91.75
news-3cl-2	**93.31**	92.98	92.06	92.89	92.39	91.68	91.39	89.38	92.64	54.98	55.64	92.98	92.18	89.72
news-3cl-3	91.18	93.03	93.24	**93.99**	93.78	91.39	91.01	81.68	90.55	64.50	57.61	93.11	93.07	90.84
news-5cl-1	86.32	**87.98**	87.57	87.80	87.47	86.02	86.50	76.40	81.04	48.72	27.73	86.90	87.30	86.52
news-5cl-2	79.48	78.25	81.83	77.80	81.68	77.23	80.88	60.41	75.28	51.88	47.60	77.25	81.41	**82.51**
news-5cl-3	73.60	81.02	81.09	80.29	80.77	79.91	78.91	61.01	76.00	41.68	27.83	80.97	80.22	**81.92**

Table 3: Classification accuracy (correct classification rate) in percent for the bag-of-paths based distances and the competing methods, obtained on each dataset, using 5 social dimensions. Only the results for graphs with 20% labeling rate are reported. The best performing method of each data set is highlighted in boldface.

Moreover, in order to rate globally the results of each method, we use a nonparametric Friedman-Nemenyi statistical test [27] allowing to compare them across all the datasets. The obtained ranking scores are presented in Figure 2 and are similar to those provided by the Borda ranking. The figure confirms that the BoP and RSP distances provide good results, although not significantly different from the logarithmic forest and the two baseline methods (the modularity matrix Q and the sum-of-similarities SoS). This is partly because the Friedman-Nemenyi test is rather conservative, especially when comparing many different techniques.

Therefore, in order to further investigate the results, we also computed some pairwise comparisons through a nonparametric one-sided Wilcoxon signed-rank test for matched data ($\alpha = 0.05$). These paired tests show that all the introduced bag-of-paths methods (BoPP-g, BoPP-mds, BoPS-g, BoPS-mds) are significantly better than our first baseline (Q, eigenvectors extracted from the modularity matrix), but not necessarily better than the second baseline (SoS, the sum-of-similarities method). Indeed, only one method, BoPS-mds, provided significantly better results than SoS (but close to the critical value, p-value = 0.033). This confirms that the SoS can be considered as a good baseline which, in addition, is simple to implement, efficient, and scales to large, sparse, networks [71]. Moreover, by examining further Table 3, SoS is the only method achieving good performance on the imdb dataset while almost all the other methods fail on this dataset.

Although a little under the bag-of-paths based approaches, note that the randomized shortest path (RSP) and the logarithmic forest (LF) methods associated to the gaussian transformation are also competitive, consistently providing good results, and significantly better than our first baseline (Q). Note also that this simple modularity matrix based method Q, although below the best methods, especially in the 5-classes setting, provides reasonable results.

Curiously, the spectral method applied to the three kernels (the Markov diffusion kernel (MD), the regularized commute-time kernel (RCT) and the regularized Laplacian kernel (RL)) provides bad results (all three kernels perform significantly worse than the two baselines). This is especially odd, as these kernels obtained good results when used in a sum-of-similarities context [34, 71] – see the results obtained by the sum-of-similarities based on the RCT kernel (SoS) in Table 3 which is not statistically different from the best method. This could be related to the recent comparison in [49] showing that taking the logarithm of some well-known kernels improves the results in node clustering tasks.

Concerning the transformation from distances to inner products of Equations (43) and (44), the Gaussian kernel often provides slightly better results than multidimensional scaling, but not always so.

In summary, these experiments showed that the introduced BoP families of distances (BoPP, BoPS), but also the already known randomized shortest path (RSP) and the logarithmic forest (LF) distances, achieve good results in comparison with our two baseline methods (Q and SoS) on the investigated datasets. However, we found that the introduced distances are not necessarily significantly better (although globally ranked better) than the second baseline, the sum-of-similarities method (SoS) based on the RCT kernel [34, 71]. Because this SoS technique is fast and scales to large graphs [71], it can be concluded that the introduced distances do not bring much added value here in our semi-supervised

Method	Rank	Score
BoPP-g	1	162
BoPS-mds	2	155
BoPP-mds	3	143
RSP-g	4	141
BoPS-g	5	140
LogF-g	6	127
BoPM	7	125
RSP-mds	8	118
SoS	9	109
LogF-mds	10	95
Q	11	89
MD	12	72
RCT	13	38
RL	14	29

Table 4: Ranking of the different classification methods according to Borda's method (the higher score, the better).

classification tasks. Still, this has to be confirmed in larger experiments. Indeed, in further work, we plan to conduct a systematic, comprehensive, comparison of families of distances and kernels on clustering, classification and dimensionality reduction tasks.

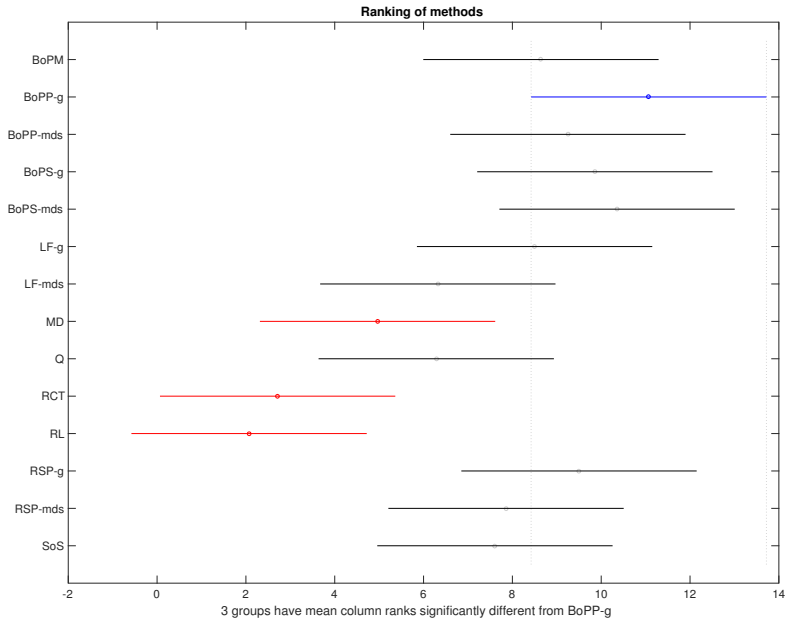


Figure 2: Friedman-Nemenyi ranking over the 14 datasets (the larger, the better). Two methods are considered as significantly different when their confidence intervals do not overlap. The best ranked method (BoPP-g) is highlighted.

8. Conclusion and further work

This work introduced the bag-of-paths framework considering a bag containing the set of paths in the network. By defining a Gibbs-Boltzmann distribution on this set of paths penalizing long paths, we were able to derive various interesting quantities such as distance measures between nodes. It is also shown that one of the two introduced distance measures has some nice properties, like interpolating between the shortest-path distance and the resistance distance (up to a constant factor). Experiments have shown that the BoP framework can provide competitive algorithms within a clear theoretical framework.

Indeed, as demonstrated in semi-supervised classification experiments, the kernels associated to the distance measures derived from the bag-of-paths probabilities achieve good results. Consistency of performance across the different datasets (except imdb) shows that the bag-of-paths framework seems to induce some promising distance and similarity measures on graphs, based on its structure.

The framework is rich and other quantities of interest can be defined, which is pursued in parallel. For instance, a betweenness measure can be defined as $\mathrm{P}(int = j|s = i, e = k)$, the probability that a path starting in i and ending in k visits j as an intermediate node [64]. Another idea is to reformulate the modularity matrix in terms of paths instead of direct links [28]. Still another application is the definition of a new robustness measure capturing the criticality of the nodes [65]. The idea then would be to compute the change in accessibility between nodes when deleting one node within the BoP framework. Nodes having a wide impact on reachability are then considered as highly critical.

Another idea would be to investigate graph cuts by considering paths instead of links. We also plan to evaluate experimentally the potential distance (see Equation (34)) as a distance between sequences of characters by adapting it to a directed acyclic graph, as in [37].

Finally, we plan to make a systematic experimental comparison between families of distances and kernels on clustering, semi-supervised classification and dimensionality reduction tasks, while trying to analyze the theoretical properties of the proposed distances families by following [43]. In particular, we will investigate the new kernels introduced recently in [49] where it is shown on node clustering tasks that taking the logarithm of well-known kernels improves significantly the performance.

Acknowledgments

This work was partially supported by the Immediate and the Brufence projects funded by InnovIris (Brussels Region), as well as former projects funded by the Walloon region. Ilkka Kivimäki was partially funded by Emil Aaltonen Foundation, Finland. We thank these institutions for giving us the opportunity to conduct both fundamental and applied research. We also thank Bertrand Lebichot, Dr. Guillaume Guex, Dr. Yutaro Shigeto and Dr. Chebotarev for helping us during the experiments and making various interesting comments. Finally, we thank the anonymous reviewers for their remarks.

Appendix

Appendix A. Sum of reference probabilities over hitting paths

In this appendix, it is shown that the sum over all hitting paths of the reference probabilities is equal to one. We thus have to compute

$$\sum_{\wp\in\mathcal{P}^{\mathrm{h}}} \tilde{\mathrm{P}}^{\mathrm{ref}}(\wp) = \sum_{t=0}^{\infty} \sum_{\wp\in\mathcal{P}^{\mathrm{h}}(t)} \tilde{\mathrm{P}}^{\mathrm{ref}}(\wp) = \sum_{t=0}^{\infty} \sum_{i,j=1}^{n} \sum_{\wp_{ij}\in\mathcal{P}_{ij}^{\mathrm{h}}(t)} \tilde{\mathrm{P}}^{\mathrm{ref}}(\wp_{ij})$$
$$= \frac{1}{n^2} \sum_{t=0}^{\infty} \sum_{i,j=1}^{n} \sum_{\wp_{ij}\in\mathcal{P}_{ij}^{\mathrm{h}}(t)} \tilde{\pi}^{\mathrm{ref}}(\wp_{ij}) \stackrel{?}{=} 1 \tag{A.1}$$

where $\mathcal{P}^{\mathrm{h}}(t)$ is the set of all hitting paths of length exactly equal to t and $\mathcal{P}_{ij}^{\mathrm{h}}(t)$ the set of such hitting paths connecting i to j. As stated before, because we assume that the a priori probability of choosing the starting node and ending node is uniform, $\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp_{ij}) = \frac{1}{n^2}\pi^{\mathrm{ref}}(\wp_{ij})$ with $\pi^{\mathrm{ref}}(\wp_{ij})$ being the likelihood of the path $\wp_{ij}$, i.e., the product of transition probabilities $p_{ll'}^{\mathrm{ref}}$ along the path of length t, $\pi^{\mathrm{ref}}(\wp_{ij}) = \prod_{\tau=1}^{t} p_{k_{\tau-1}k_\tau}^{\mathrm{ref}}$ with $k_0 = i$, $k_t = j$ and no intermediate node being equal to node j.

As we are concerned with hitting paths stopping in node j, let us consider the absorbing, killing, Markov chain on G with transition probabilities $p_{ll'}^{\mathrm{ref}}$ for $l \neq j$ and $p_{jl'}^{\mathrm{ref}} = 0$ for all l'. In other words, node j is made killing and absorbing.

We now introduce a new quantity, $q_k^{(ij)}(t)$, on this absorbing Markov chain, defined as the probability of finding the process in state k at time t when considering walks from starting node i to absorbing node j. This probability can easily be computed thanks to the following recurrence relation

$$\begin{cases} q_k^{(ij)}(0) = \delta_{ik} & \text{for } t = 0 \\ q_k^{(ij)}(t) = \sum\limits_{\substack{l=1 \\ l\neq j}}^{n} q_l^{(ij)}(t-1)\, p_{lk}^{\mathrm{ref}} & \text{for } t \geq 1 \end{cases} \tag{A.2}$$

which says that the probability of being in node k at time t is the sum of the probabilities of being in any node l (except node j which is absorbing) at time $t-1$ times the probability of jumping from l to k. When $k = j$, the quantity computes the probability of being absorbed in node j at time t, given that we started from i at time 0.

Let us now examine the last quantity appearing in Equation (A.1), the sum of hitting paths likelihoods from i to j, assuming $i \neq j$,

$$\sum_{\wp_{ij}\in\mathcal{P}_{ij}^{\mathrm{h}}(t)} \tilde{\pi}^{\mathrm{ref}}(\wp_{ij}) = \sum_{\substack{k_1=1 \\ k_1\neq j}}^{n} \sum_{\substack{k_2=1 \\ k_2\neq j}}^{n} \cdots \sum_{\substack{k_{t-1}=1 \\ k_{t-1}\neq j}}^{n} p_{ik_1}^{\mathrm{ref}} p_{k_1k_2}^{\mathrm{ref}} p_{k_2k_3}^{\mathrm{ref}} \cdots p_{k_{t-1}j}^{\mathrm{ref}} \quad \text{for } t > 0 \tag{A.3}$$

and it is equal to 0 when $t = 0$ because there is no path of length zero connecting two different nodes.

But the second-hand quantity in this last equation is nothing else than the sequential application of recurrence (A.2) for $t, t-1, \ldots, 0$, therefore computing

$q_j^{(ij)}(t)$, that is, the probability of being absorbed in node j in exactly t steps. Therefore, $\sum_{\wp_{ij}\in\mathcal{P}_{ij}^{\mathrm{h}}(t)} \tilde{\pi}^{\mathrm{ref}}(\wp_{ij}) = q_j^{(ij)}(t)$ when $i \neq j$.

Moreover, as we know that the process necessarily ends in absorbing node j at some point (see, e.g., [39]), $\sum_{t=0}^{\infty} q_j^{(ij)}(t) = 1$ holds when $i \neq j$.

Conversely, when $i = j$, the probability of finding the process in node j is 1 at $t = 0$ (a zero-length path) and then collapses to 0 when $t > 0$, which also provides $\sum_{t=0}^{\infty} q_j^{(jj)}(t) = 1$.

Equation (A.1) then becomes

$$\sum_{\wp\in\mathcal{P}^{\mathrm{h}}} \tilde{\mathrm{P}}^{\mathrm{ref}}(\wp) = \frac{1}{n^2}\sum_{i,j=1}^{n}\sum_{t=0}^{\infty} q_j^{(ij)}(t) = \frac{1}{n^2}\sum_{i,j=1}^{n} 1 = 1 \tag{A.4}$$

which is the desired result. In addition, this also shows that

$$\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp) = 1, \tag{A.5}$$

that is, the sum over the path likelihoods is equal to 1 for hitting paths.

Appendix B. Computation of the entries of $\mathbf{Z}^{(-\mathrm{j})}$ in terms of the fundamental matrix

All the entries of $\mathbf{Z}^{(-j)}$ can be computed efficiently in terms of the fundamental matrix $\mathbf{Z} = (\mathbf{I} - \mathbf{W})^{-1}$.

This result can be understood as follows. Each non-hitting path $\wp_{ij} \in \mathcal{P}_{ij}$ can be split uniquely into two sub-paths, before hitting node j for the first time, $\wp_{ij}^{\mathrm{h}} \in \mathcal{P}_{ij}^{\mathrm{h}}$, and after hitting node j, $\wp_{jj} \in \mathcal{P}_{jj}$. These two sub-paths can be chosen independently because their concatenation is a valid path, with $\wp_{ij}^{\mathrm{h}} \circ \wp_{jj} \in \mathcal{P}_{ij}$ being the concatenation of the two paths. Now, as $\tilde{c}(\wp_{ij}) = \tilde{c}(\wp_{ij}^{\mathrm{h}}) + \tilde{c}(\wp_{jj})$ and $\tilde{\pi}^{\mathrm{ref}}(\wp_{ij}) = \tilde{\pi}^{\mathrm{ref}}(\wp_{ij}^{\mathrm{h}})\tilde{\pi}^{\mathrm{ref}}(\wp_{jj})$ for any $\wp_{ij} = \wp_{ij}^{\mathrm{h}} \circ \wp_{jj}$, we obtain

$$\begin{aligned}
z_{ij} &= \sum_{\wp_{ij}\in\mathcal{P}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp_{ij}) \exp[-\theta\tilde{c}(\wp_{ij})] \\
&= \sum_{\substack{\wp_{ij}^{\mathrm{h}}\in\mathcal{P}_{ij}^{\mathrm{h}} \\ \wp_{jj}\in\mathcal{P}_{jj}}} \tilde{\pi}^{\mathrm{ref}}(\wp_{ij}^{\mathrm{h}})\tilde{\pi}^{\mathrm{ref}}(\wp_{jj}) \exp[-\theta\tilde{c}(\wp_{ij}^{\mathrm{h}})] \exp[-\theta\tilde{c}(\wp_{jj})] \\
&= \left(\sum_{\wp_{ij}^{\mathrm{h}}\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp_{ij}^{\mathrm{h}}) \exp[-\theta\tilde{c}(\wp_{ij}^{\mathrm{h}})]\right)\left(\sum_{\wp_{jj}\in\mathcal{P}_{jj}} \tilde{\pi}^{\mathrm{ref}}(\wp_{jj}) \exp[-\theta\tilde{c}(\wp_{jj})]\right) \\
&= z_{ij}^{(-j)} z_{jj}
\end{aligned} \tag{B.1}$$

and therefore $z_{ij}^{(-j)} = z_{ij}/z_{jj}$. Using this result, Equation (22) can be developed as

$$\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta\tilde{c}(\wp)] = z_{ij}^{(-j)} = \frac{z_{ij}}{z_{jj}} \tag{B.2}$$

Appendix C. Triangle inequality proof for the surprisal distance

In order for Δ_{ij}^{sur} to be a distance measure, it has to be shown that it obeys the triangle inequality, $\Delta_{ik}^{\text{sur}} \leq \Delta_{ij}^{\text{sur}} + \Delta_{jk}^{\text{sur}}$ for all i, j, k. Note that $\Delta_{ij}^{\text{sur}} = \infty$ when node i and node j are not connected (they belong to different connected components) – this is why we require G to be strongly connected. In addition, note that the triangle inequality is trivially satisfied if either $i = j$, $j = k$ or $i = k$. Thus, we only need to prove the case $i \neq j \neq k \neq i$.

In order to prove the result, consider the set of paths $\mathcal{P}_{ik}$ from node i to node k. We now compute the probability that such paths pass through an *intermediate* node $int = j$ where $i \neq j \neq k \neq i$,

$$\mathrm{P}(s = i, int = j, e = k) = \frac{\sum_{\wp \in \mathcal{P}_{ik}} \delta(j \in \wp)\, \tilde{\pi}^{\text{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)]}{\sum_{\wp' \in \mathcal{P}} \tilde{\pi}^{\text{ref}}(\wp') \exp[-\theta \tilde{c}(\wp')]} \tag{C.1}$$

where $\delta(j \in \wp)$ is a Kronecker delta equal to 1 if the path $\wp$ contains (at least once) node j, and 0 otherwise. It is clear from Equations (21) and (C.1) that

$$\mathrm{P}(s = i, e = k) \geq \mathrm{P}(s = i, int = j, e = k) \quad \text{for } i \neq j \neq k \neq i \tag{C.2}$$

Let us transform Equation (C.1), using the fact that each path $\wp_{ik}$ between i and k passing through j can be decomposed uniquely into a *hitting* sub-path $\wp_{ij}$ from i to j and a non-hitting sub-path $\wp_{jk}$ from j to k. The sub-path $\wp_{ij}$ is found by following path $\wp_{ik}$ until reaching j for the first time. Therefore, for $i \neq j \neq k \neq i$,

$$\begin{aligned}
\mathrm{P}(s = i, int = j, e = k) &= \frac{\sum_{\wp \in \mathcal{P}_{ik}} \delta(j \in \wp)\, \tilde{\pi}^{\text{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)]}{\mathcal{Z}} \\
&= \frac{\sum_{\wp_{ij} \in \mathcal{P}_{ij}^{\text{h}}} \sum_{\wp_{jk} \in \mathcal{P}_{jk}} \tilde{\pi}^{\text{ref}}(\wp_{ij}) \tilde{\pi}^{\text{ref}}(\wp_{jk}) \exp[-\theta(\tilde{c}(\wp_{ij}) + \tilde{c}(\wp_{jk}))]}{\mathcal{Z}} \\
&= \frac{\left[\sum_{\wp_{ij} \in \mathcal{P}_{ij}^{\text{h}}} \tilde{\pi}^{\text{ref}}(\wp_{ij}) \exp[-\theta \tilde{c}(\wp_{ij})]\right] \left[\sum_{\wp_{jk} \in \mathcal{P}_{jk}} \tilde{\pi}^{\text{ref}}(\wp_{jk}) \exp[-\theta \tilde{c}(\wp_{jk})]\right]}{\mathcal{Z}} \\
&= \mathcal{Z}_{\text{h}} \frac{\left[\sum_{\wp_{ij} \in \mathcal{P}_{ij}^{\text{h}}} \tilde{\pi}^{\text{ref}}(\wp_{ij}) \exp[-\theta \tilde{c}(\wp_{ij})]\right]}{\mathcal{Z}_{\text{h}}} \frac{\left[\sum_{\wp_{jk} \in \mathcal{P}_{jk}} \tilde{\pi}^{\text{ref}}(\wp_{jk}) \exp[-\theta \tilde{c}(\wp_{jk})]\right]}{\mathcal{Z}} \\
&= \mathcal{Z}_{\text{h}}\, \mathrm{P}_{\text{h}}(s = i, e = j)\, \mathrm{P}(s = j, e = k), \text{ for } i \neq j \neq k \neq i
\end{aligned} \tag{C.3}$$

Combining Inequality (C.2) and Equation (C.3) yields

$$\mathrm{P}(s = i, e = k) \geq \mathcal{Z}_{\text{h}}\, \mathrm{P}_{\text{h}}(s = i, e = j)\, \mathrm{P}(s = j, e = k), \text{ for } i \neq j \neq k \neq i \tag{C.4}$$

Replacing the non-hitting bag-of-paths probabilities by their expressions (see Equation (18)) in function of the elements of the fundamental matrix, $\mathrm{P}(s =$

$i, e = k) = z_{ik}/\mathcal{Z}$ and $\mathrm{P}(s = j, e = k) = z_{jk}/\mathcal{Z}$, in the previous Inequality (C.4) provides $z_{ik}/\mathcal{Z} \geq \mathcal{Z}_\mathrm{h}\,\mathrm{P}_\mathrm{h}(s = i, e = j)\, z_{jk}/\mathcal{Z}$. Further dividing each member by $(\mathcal{Z}_\mathrm{h} z_{kk})$ gives $z_{ik}/(\mathcal{Z}_\mathrm{h} z_{kk}) \geq \mathcal{Z}_\mathrm{h}\,\mathrm{P}_\mathrm{h}(s = i, e = j)\, z_{jk}/(\mathcal{Z}_\mathrm{h} z_{kk})$. Finally, using $\mathrm{P}_\mathrm{h}(s = i, e = k) = z_{ik}/(\mathcal{Z}_\mathrm{h} z_{kk})$ (see Equation (25)), we obtain

$$\mathrm{P}_\mathrm{h}(s = i, e = k) \geq \mathcal{Z}_\mathrm{h}\,\mathrm{P}_\mathrm{h}(s = i, e = j)\,\mathrm{P}_\mathrm{h}(s = j, e = k) \tag{C.5}$$

for $i \neq j \neq k \neq i$. Now, from Equation (26) and the fact that the z_{ij} are nonnegative, it is clear that $\mathcal{Z}_\mathrm{h} \geq 1$; thus

$$\mathrm{P}_\mathrm{h}(s = i, e = k) \geq \mathrm{P}_\mathrm{h}(s = i, e = j)\,\mathrm{P}_\mathrm{h}(s = j, e = k), \text{ for } i \neq j \neq k \neq i \tag{C.6}$$

Finally, by taking $-\log$ of Inequality (C.6),

$$-\log \mathrm{P}_\mathrm{h}(s = i, e = k) \leq -\log \mathrm{P}_\mathrm{h}(s = i, e = j) - \log \mathrm{P}_\mathrm{h}(s = j, e = k), \tag{C.7}$$

for $i \neq j \neq k \neq i$. Thus, the (directed) surprisal measure, $-\log \mathrm{P}_\mathrm{h}(s = i, e = j)$, obeys the triangle inequality. Therefore the surprisal distance $\Delta_{ij}^{\mathrm{sur}} = -(\log \mathrm{P}_\mathrm{h}(s = i, e = j) + \log \mathrm{P}_\mathrm{h}(s = j, e = i))/2$ also enjoys this property.

Appendix D. Proof of the cutpoint additive property of the potential distance

From the definition of the bag-of-paths probability (Equation (9)), as well as Equation (C.1) defining $\mathrm{P}(s = i, int = j, e = k)$, we have for $i \neq j \neq k \neq i$

$$\begin{aligned}
\mathrm{P}(s = i, e = k) &= \frac{\sum_{\wp \in \mathcal{P}_{ik}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)]}{\mathcal{Z}} \\
&= \frac{\sum_{\wp \in \mathcal{P}_{ik}} \delta(j \in \wp)\, \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)]}{\mathcal{Z}} + \frac{\sum_{\wp \in \mathcal{P}_{ik}} (1 - \delta(j \in \wp))\, \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)]}{\mathcal{Z}} \\
&= \mathrm{P}(s = i, int = j, e = k) + \frac{\sum_{\wp \in \mathcal{P}_{ik}} \delta(j \notin \wp)\, \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)]}{\mathcal{Z}}
\end{aligned} \tag{D.1}$$

Now, substituting $\mathrm{P}(s = i, int = j, e = k)$ by $\mathcal{Z}_\mathrm{h}\,\mathrm{P}_\mathrm{h}(s = i, e = j)\mathrm{P}(s = j, e = k)$ (see Equation (C.3)) in the previous equation yields

$$\begin{aligned}
\mathrm{P}(s = i, e = k) =& \mathcal{Z}_\mathrm{h}\,\mathrm{P}_\mathrm{h}(s = i, e = j)\mathrm{P}(s = j, e = k) \\
&+ \frac{\sum_{\wp \in \mathcal{P}_{ik}} \delta(j \notin \wp)\, \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)]}{\mathcal{Z}}
\end{aligned} \tag{D.2}$$

Further recalling that $\mathrm{P}(s = i, e = k) = z_{ik}/\mathcal{Z}$ (Equation (18)) and $\mathrm{P}_\mathrm{h}(s = i, e = j) = z_{ij}^{\mathrm{h}}/\mathcal{Z}_\mathrm{h}$ (Equation (25)), we transform Equation (D.2) into

$$z_{ik} = z_{ij}^{\mathrm{h}} z_{jk} + \sum_{\wp \in \mathcal{P}_{ik}} \delta(j \notin \wp)\, \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)] \tag{D.3}$$

Dividing both sides of the previous equation by z_{kk} and recalling that $z^{\mathrm{h}}_{ik} = z_{ik}/z_{kk}$ (Equation (24)) provides

$$z^{\mathrm{h}}_{ik} = z^{\mathrm{h}}_{ij} z^{\mathrm{h}}_{jk} + \frac{1}{z_{kk}} \sum_{\wp \in \mathcal{P}_{ik}} \delta(j \notin \wp)\, \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)] \tag{D.4}$$

and we recover $z^{\mathrm{h}}_{ik} \geq z^{\mathrm{h}}_{ij} z^{\mathrm{h}}_{jk}$ (Equation (31)). The equality $z^{\mathrm{h}}_{ik} = z^{\mathrm{h}}_{ij} z^{\mathrm{h}}_{jk}$ ($i \neq j \neq k \neq i$) holds if and only if $\sum_{\wp \in \mathcal{P}_{ik}} \delta(j \notin \wp)\, \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)] = 0$, which only occurs when all paths connecting i and k visit node j. Thus, it is clear that $\Delta^{\phi}_{ik} = \Delta^{\phi}_{ij} + \Delta^{\phi}_{jk}, i \neq j \neq k \neq i$ if and only if all paths $\wp \in \mathcal{P}^{\mathrm{h}}_{ik}$ and $\wp \in \mathcal{P}^{\mathrm{h}}_{ki}$ connecting node i and node k pass through node j. This property is called the cutpoint additivity or graph-geodetic property in [15].

Appendix E. Asymptotic result: for an undirected graph, the Δ^{ϕ} distance converges to the shortest-path distance when $\theta \to \infty$

There are two ways to prove this property, each of them having its own benefits. The first proof is based on the bag-of-paths framework and is shorter. The second proof is inspired by [98] and is longer, but establishes some interesting links with the Bellman-Ford formula for computing the shortest-path distance in a network (see, e.g., [7, 20, 25, 85, 92]).

Appendix E.1. First proof

Assuming $i \neq j$ and $\theta > 0$, let us recall (Equation (33)), that is, $\Delta^{\phi}_{ij} = (\phi(i,j) + \phi(j,i))/2$ with $\phi(i,j) = -\frac{1}{\theta} \log z^{\mathrm{h}}_{ij}$, and where z^{h}_{ij} is given by (Equation (24), recalled here for convenience):

$$z^{\mathrm{h}}_{ij} = \sum_{\wp \in \mathcal{P}^{\mathrm{h}}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)] \tag{E.1}$$

which is always positive for a strongly connected graph.

We now have to compute the asymptotic form of z^{h}_{ij} for $\theta \to \infty$ or, equivalently, $T \to 0$. Let the lowest-cost (shortest) paths from i to j be denoted as $\{\wp^*_k\}$ and let $c^* = \tilde{c}(\wp^*_k)$ be the cost of such a lowest-cost path. c^* is therefore the minimum cost among all possible paths from i to j. Say there are $m \geq 1$ such lowest-cost paths. Now, as $\sum_{\wp \in \mathcal{P}^{\mathrm{h}}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp) = 1$ (see Equation A.5), it is clear that z^{h}_{ij} is bounded by

$$z^{\mathrm{h}}_{ij} \leq \sum_{\wp \in \mathcal{P}^{\mathrm{h}}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta c^*] = \exp[-\theta c^*] \sum_{\wp \in \mathcal{P}^{\mathrm{h}}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp) = \exp[-\theta c^*] \tag{E.2}$$

and is therefore finite. We also observe that it converges exponentially to 0 when $\theta \to \infty$. Moreover, this last inequality implies

$$\sum_{\wp \in \mathcal{P}^{\mathrm{h}}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta(\tilde{c}(\wp) - c^*)] \leq 1 \tag{E.3}$$

which shows that the quantity on the left-hand side is bounded.

We can now rewrite

$$z_{ij}^{\mathrm{h}} = \sum_{\wp \in \mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta \tilde{c}(\wp)] = \exp[-\theta c^*] \sum_{\wp \in \mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta(\tilde{c}(\wp) - c^*)]$$

$$= \exp[-\theta c^*] \left(\sum_{i=1}^{m} \tilde{\pi}^{\mathrm{ref}}(\wp_i^*) + \sum_{\substack{\wp \in \mathcal{P}_{ij}^{\mathrm{h}} \\ \tilde{c}(\wp) > c^*}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta(\tilde{c}(\wp) - c^*)] \right) \tag{E.4}$$

Let us now compute the potential $\phi(i,j) = -\frac{1}{\theta} \log z_{ij}^{\mathrm{h}}$ when $\theta \to \infty$. Using Equation (E.4), we get

$$\begin{aligned}
\phi(i,j) &= -\frac{1}{\theta} \log z_{ij}^{\mathrm{h}} \\
&= -\frac{1}{\theta} \log \left[\exp[-\theta c^*] \left(\sum_{i=1}^{m} \tilde{\pi}^{\mathrm{ref}}(\wp_i^*) + \sum_{\substack{\wp \in \mathcal{P}_{ij}^{\mathrm{h}} \\ \tilde{c}(\wp) > c^*}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta(\tilde{c}(\wp) - c^*)] \right) \right] \\
&= c^* - \frac{1}{\theta} \log \left(\sum_{i=1}^{m} \tilde{\pi}^{\mathrm{ref}}(\wp_i^*) + \sum_{\substack{\wp \in \mathcal{P}_{ij}^{\mathrm{h}} \\ \tilde{c}(\wp) > c^*}} \tilde{\pi}^{\mathrm{ref}}(\wp) \exp[-\theta(\tilde{c}(\wp) - c^*)] \right) \\
&\xrightarrow{\theta \to \infty} c^*
\end{aligned} \tag{E.5}$$

Here, the last limit applies because, following Equation (E.3), the expression inside the logarithm is finite and strictly positive (the first term is a positive value and the second is positive and bounded (see Equation (E.3))).

Moreover, observing that, in the case of an undirected graph, the lowest cost from j to i is equal to the lowest cost from i to j (i.e., c^*), the distance $\Delta_{ij}^{\phi} = \frac{\phi(i,j)+\phi(j,i)}{2} \xrightarrow{\theta \to \infty} c^*$. Therefore, the bag-of-hitting-paths potential distance provides the shortest-path distance when $\theta \to \infty$ for undirected graphs.

Appendix E.2. Second proof

The second proof starts from Equation (29), where we replace $w_{ij} = p_{ij}^{\mathrm{ref}} \exp[-\theta c_{ij}]$ in this expression with node k absorbing,

$$z_{ik}^{\mathrm{h}} = \begin{cases} \sum_{j=1}^{n} p_{ij}^{\mathrm{ref}} \exp[-\theta c_{ij}] \, z_{jk}^{\mathrm{h}} & \text{for } i \neq k \\ 1 & \text{for } i = k \text{ (boundary condition)} \end{cases} \tag{E.6}$$

Let us now compute the value of the potential $\phi(i,k)$ (Equation (33)) for

$i \neq k$ (when $i = k$, $\phi(k,k) = -\frac{1}{\theta}\log(z_{kk}/z_{kk}) = 0$),

$$\begin{aligned}
\phi(i,k) &= -\frac{1}{\theta}\log z^{\mathrm{h}}_{ik} = -\frac{1}{\theta}\log\left[\sum_{j=1}^{n} p^{\mathrm{ref}}_{ij}\exp[-\theta c_{ij}]\, z^{\mathrm{h}}_{jk}\right] \\
&= -\frac{1}{\theta}\log\left[\sum_{j=1}^{n} p^{\mathrm{ref}}_{ij}\exp[-\theta c_{ij}]\,\exp[-\theta(-\frac{1}{\theta}\log z^{\mathrm{h}}_{jk})]\right] \\
&= -\frac{1}{\theta}\log\left[\sum_{j=1}^{n} p^{\mathrm{ref}}_{ij}\exp[-\theta c_{ij}]\,\exp[-\theta\phi(j,k)]\right] \\
&= -\frac{1}{\theta}\log\left[\sum_{j\in\mathcal{S}ucc(i)} p^{\mathrm{ref}}_{ij}\exp[-\theta(c_{ij}+\phi(j,k))]\right] \qquad \text{(E.7)}
\end{aligned}$$

which provides a recurrence formula for computing $\phi(i,k)$, together with the boundary condition $\phi(k,k) = 0$.

Let us now study the behavior of this equation when $\theta \to \infty$. We first observe that both the numerator and the denominator tend to $+\infty$. Now, in order to simplify the notations, we will study the softmin function [24, 98], $\mathrm{softmin}_{\mathbf{q},\theta}(\mathbf{x}) = -\log(\sum_{j=1}^{n} q_j \exp[-\theta x_j])/\theta$ with $\sum_{j=1}^{n} q_j = 1$ and all $q_j \geq 0$ instead, where we define $x_j = (c_{ij} + \phi(j,k))$ and $q_j = p^{\mathrm{ref}}_{ij}$ (the development is inspired by [98]). Let us further define $x^* = \min_j(x_j)$ so that $(x_j - x^*) \geq 0$; we then have

$$\begin{aligned}
\lim_{\theta\to\infty} \mathrm{softmin}_{\mathbf{q},\theta}(\mathbf{x}) &= \lim_{\theta\to\infty} -\frac{\log\left[\sum_{j=1}^{n} q_j \exp[-\theta x_j]\right]}{\theta} \\
&= \lim_{\theta\to\infty} -\frac{\log\left[\exp[-\theta x^*]\sum_{j=1}^{n} q_j \exp[-\theta(x_j - x^*)]\right]}{\theta} \\
&= \lim_{\theta\to\infty}\left[x^* - \frac{\log\left[\sum_{j=1}^{n} q_j \exp[-\theta(x_j - x^*)]\right]}{\theta}\right] \\
&= x^* - \lim_{\theta\to\infty} \frac{\log\left[\sum_{j=1}^{n} q_j \exp[-\theta(x_j - x^*)]\right]}{\theta} \\
&= x^* \qquad \text{(E.8)}
\end{aligned}$$

and the last limit is 0 because no term in the exponential is positive and at least one of the x_j is *exactly* equal to x^* (the minimum) so that the sum $\sum_{j=1}^{n} q_j \exp[-\theta(x_j - x^*)]$ is non-zero, and thus strictly positive.

Thus, when $\theta \to \infty$, Equation (E.7) becomes $\phi(i,k) = \min_j(c_{ij} + \phi(j,k))$ for $i \neq k$ and $\phi(k,k) = 0$ which is the well-known Bellman-Ford formula for computing the shortest-path distance in an undirected graph (see, e.g., [7, 20, 25, 52, 85, 92]). Moreover, for such an undirected graph, the shortest path from i to j is equal to the shortest path from j to i, which implies that Δ^ϕ reduces to the shortest-path distance too when $\theta \to \infty$.

Appendix F. Asymptotic result: for an undirected graph, the Δ^ϕ distance converges to half the commute cost distance when $\theta \to 0^+$

Let us show that the Δ^ϕ distance is half the commute-cost distance when $\theta \to 0^+$. As before, there are two ways to prove this property. The first proof is based on the bag-of-paths framework and is somewhat shorter. The second proof, also inspired by [98], establishes some interesting links with the Bellman-Ford recurrence formula [56, 79, 86, 102].

Appendix F.1. First proof

From Equations (33) and (E.1),

$$\begin{aligned}\Delta_{ij}^{\phi} &= -\frac{(\log z_{ij}^{\mathrm{h}} + \log z_{ji}^{\mathrm{h}})}{2\theta} \\ &= -\frac{\log(\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]) + \log(\sum_{\wp\in\mathcal{P}_{ji}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)])}{2\theta}\end{aligned} \tag{F.1}$$

and, because $\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp) = 1$ for hitting paths (see Equation A.5), both the numerator and the denominator tend to zero when $\theta \to 0^+$. For taking the limit $\theta \to 0^+$ of the whole expression (F.1), we apply l'Hospital's rule (taking the derivative of the numerator and the denominator with respect to θ and then the limit $\lim_{\theta\to 0^+}$ of the resulting expression). Because the Gibbs-Boltzmann probability distribution over the hitting paths tends to $\tilde{\pi}^{\mathrm{ref}}$ when $\theta \to 0^+$ (see Equation (3)), this provides

$$\lim_{\theta\to 0^+} \Delta_{ij}^{\phi} = \frac{\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\,\tilde{c}(\wp) + \sum_{\wp\in\mathcal{P}_{ji}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\,\tilde{c}(\wp)}{2} \tag{F.2}$$

The quantity $\sum_{\wp\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp)\,\tilde{c}(\wp)$ can be interpreted as the average first-passage cost from i to j, i.e. the average cost undergone by a random walker using transition probabilities p_{ij}^{ref} for reaching destination node j for the first time when starting from i. Consequently, the average of the two quantities defined in (F.2) is half the commute-cost distance.

Appendix F.2. Second proof

Restarting from Equation (E.7), we now have to take the limit $\theta \to 0^+$. Assuming $\sum_{j=1}^n q_j = 1$, let us compute the limit $\theta \to 0^+$ of $\mathrm{softmin}_{\mathbf{q},\theta}(\mathbf{x})$,

instead of $\theta \to \infty$ in Equation (E.8), and apply as before l'Hospital's rule

$$\lim_{\theta\to 0^+} \operatorname{softmin}_{\mathbf{q},\theta}(\mathbf{x}) = \lim_{\theta\to 0^+} -\frac{\log\left(\sum_{j=1}^{n} q_j \exp[-\theta x_j]\right)}{\theta}$$
$$= \lim_{\theta\to 0^+} \frac{\sum_{j=1}^{n} q_j x_j \exp[-\theta x_j]}{\sum_{j'=1}^{n} q_{j'} \exp[-\theta x_{j'}]} = \frac{\sum_{j=1}^{n} q_j x_j}{\sum_{j'=1}^{n} q_{j'}} = \sum_{j=1}^{n} q_j x_j \quad \text{(F.3)}$$

Therefore, as in our case $x_j = (c_{ij}+\phi(j,k))$ and $q_j = p_{ij}^{\mathrm{ref}}$ with $\sum_{j=1}^{n} p_{ij}^{\mathrm{ref}} = 1$, we obtain $\phi(i,k) = \sum_{j=1}^{n} p_{ij}^{\mathrm{ref}}(c_{ij}+\phi(j,k))$ for $i \neq k$, together with the boundary condition $\phi(k,k) = 0$. But this is exactly the recurrence formula computing the average first-passage cost in a regular Markov chain [56, 79, 86, 102]. Thus, when $\theta \to 0^+$, $\Delta^{\phi} = (\phi(i,j) + \phi(j,i))/2$ reduces to half the commute-cost distance between i and j.

Appendix G. Triangle inequality for hitting paths and weighted nodes

To prove the result we simply adapt the corresponding proof of Appendix C. Note that Equation (C.2) still holds. Moreover, Equation (C.3) becomes

$$\mathrm{P}(s=i, int=j, e=k) = \frac{q_i^{\mathrm{s}} q_k^{\mathrm{e}} \sum_{\wp\in\mathcal{P}_{ik}} \delta(j\in\wp)\,\tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\mathcal{Z}_{\mathrm{w}}}$$
$$= \frac{q_i^{\mathrm{s}} q_k^{\mathrm{e}} \sum_{\wp_{ij}\in\mathcal{P}_{ij}^{\mathrm{h}}} \sum_{\wp_{jk}\in\mathcal{P}_{jk}} \tilde{\pi}^{\mathrm{ref}}(\wp_{ij})\tilde{\pi}^{\mathrm{ref}}(\wp_{jk})\exp[-\theta(\tilde{c}(\wp_{ij})+\tilde{c}(\wp_{jk}))]}{\mathcal{Z}_{\mathrm{w}}}$$
$$= \frac{q_i^{\mathrm{s}}\left[\sum_{\wp_{ij}\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp_{ij})\exp[-\theta\tilde{c}(\wp_{ij})]\right] q_j^{\mathrm{e}} \times q_j^{\mathrm{s}}\left[\sum_{\wp_{jk}\in\mathcal{P}_{jk}} \tilde{\pi}^{\mathrm{ref}}(\wp_{jk})\exp[-\theta\tilde{c}(\wp_{jk})]\right] q_k^{\mathrm{e}}}{q_j^{\mathrm{s}} q_j^{\mathrm{e}} \mathcal{Z}_{\mathrm{w}}}$$
$$= \frac{\mathcal{Z}_{\mathrm{hw}}}{q_j^{\mathrm{s}} q_j^{\mathrm{e}}} \frac{\left[q_i^{\mathrm{s}} \sum_{\wp_{ij}\in\mathcal{P}_{ij}^{\mathrm{h}}} \tilde{\pi}^{\mathrm{ref}}(\wp_{ij})\exp[-\theta\tilde{c}(\wp_{ij})]\, q_j^{\mathrm{e}}\right]}{\mathcal{Z}_{\mathrm{hw}}} \frac{\left[q_j^{\mathrm{s}} \sum_{\wp_{jk}\in\mathcal{P}_{jk}} \tilde{\pi}^{\mathrm{ref}}(\wp_{jk})\exp[-\theta\tilde{c}(\wp_{jk})]\, q_k^{\mathrm{e}}\right]}{\mathcal{Z}_{\mathrm{w}}}$$
$$= \frac{\mathcal{Z}_{\mathrm{hw}}}{q_j^{\mathrm{s}} q_j^{\mathrm{e}}} \mathrm{P}_{\mathrm{h}}(s=i, e=j)\,\mathrm{P}(s=j, e=k), \text{ for } i\neq j\neq k\neq i \quad \text{(G.1)}$$

where $\mathcal{Z}_{\mathrm{w}} = \sum_{i,j=1}^{n} q_i^{\mathrm{s}} z_{ij} q_j^{\mathrm{e}}$ is the partition function for non-hitting paths (the counterpart of Equation (39) for non-hitting paths).

As for Equation (C.4), combining this last result with (C.2) yields

$$\mathrm{P}(s=i, e=k) \geq \frac{\mathcal{Z}_{\mathrm{hw}}}{q_j^{\mathrm{s}} q_j^{\mathrm{e}}} \mathrm{P}_{\mathrm{h}}(s=i, e=j)\,\mathrm{P}(s=j, e=k), \text{ for } i\neq j\neq k\neq i \quad \text{(G.2)}$$

Then, by further considering that, from Equation (39), the following inequality holds

$$\frac{\mathcal{Z}_{\mathrm{hw}}}{q_j^{\mathrm{s}} q_j^{\mathrm{e}}} = \frac{1}{q_j^{\mathrm{s}} q_j^{\mathrm{e}}} \sum_{i,k=1}^{n} q_i^{\mathrm{s}} \left(\frac{z_{ik}}{z_{kk}} \right) q_k^{\mathrm{e}} \geq 1 \tag{G.3}$$

because the term $i = k = j$ in the double sum is equal to 1 and all terms are non-negative.

We deduce that $\mathrm{P}(s = i, e = k) \geq \mathrm{P}_{\mathrm{h}}(s = i, e = j)\,\mathrm{P}(s = j, e = k)$. Then, dividing both sides by $(z_{kk}\mathcal{Z}_{\mathrm{hw}})$ and using $\mathrm{P}(s = i, e = k) = (q_i^{\mathrm{s}} z_{ik} q_k^{\mathrm{e}})/\mathcal{Z}_{\mathrm{w}}$ for weighted nodes and non-hitting paths, provides $(q_i^{\mathrm{s}} z_{ik} q_k^{\mathrm{e}})/(z_{kk}\mathcal{Z}_{\mathrm{hw}}) \geq \mathrm{P}_{\mathrm{h}}(s = i, e = j)\,(q_j^{\mathrm{s}} z_{jk} q_k^{\mathrm{e}})/(z_{kk}\mathcal{Z}_{\mathrm{hw}})$. Then, from the expression of $\mathrm{P}_{\mathrm{h}}(s = i, e = k)$ in Equation (38) and by taking $-\log$ of both sides of the inequality, we have

$$-\log \mathrm{P}_{\mathrm{h}}(s = i, e = k) \leq -\log \mathrm{P}_{\mathrm{h}}(s = i, e = j) - \log \mathrm{P}_{\mathrm{h}}(s = j, e = k) \tag{G.4}$$

which shows the triangle inequality for the directed surprisal distance and, hence, the surprisal distance, in the case of weighted nodes.

The same triangle inequality result holds for the directed potential distance with weighted nodes, defined as $\phi(i, j) \triangleq -\frac{1}{\theta} \log(q_i^{\mathrm{s}} z_{ij}^{\mathrm{h}} q_j^{\mathrm{e}})$, and z_{ij}^{h} given in Equation (40). Indeed, by replacing $\mathrm{P}(\cdot)$ and $\mathrm{P}_{\mathrm{h}}(\cdot)$ by their expressions in function of the z_{ij}^{h} in Equation (G.2) and dividing both sides by z_{kk} provides

$$q_i^{\mathrm{s}} z_{ik}^{\mathrm{h}} q_k^{\mathrm{e}} \geq \frac{1}{q_j^{\mathrm{s}} q_j^{\mathrm{e}}} \left(q_i^{\mathrm{s}} z_{ij}^{\mathrm{h}} q_j^{\mathrm{e}}\right) \left(q_j^{\mathrm{s}} z_{jk}^{\mathrm{h}} q_k^{\mathrm{e}}\right) \tag{G.5}$$

Then, because $1/q_j^{\mathrm{s}} q_j^{\mathrm{e}} \geq 1$ for every j, we obtain after taking $-\frac{1}{\theta} \log$ of both sides

$$-\frac{1}{\theta} \log(q_i^{\mathrm{s}} z_{ik}^{\mathrm{h}} q_k^{\mathrm{e}}) \leq -\frac{1}{\theta} \log(q_i^{\mathrm{s}} z_{ij}^{\mathrm{h}} q_j^{\mathrm{e}}) - \frac{1}{\theta} \log(q_j^{\mathrm{s}} z_{jk}^{\mathrm{h}} q_k^{\mathrm{e}}) \tag{G.6}$$

which proves triangle inequality for the directed potential distance, and therefore also for the potential distance with priors on nodes.

References

[1] S. Abney. *Semisupervised learning for computational linguistics.* Chapman and Hall/CRC, 2008.

[2] T. Akamatsu. Cyclic flows, markov process and stochastic traffic assignment. *Transportation Research B*, 30(5):369–386, 1996.

[3] M. Alamgir and U. von Luxburg. Phase transition in the family of p-resistances. In *Advances in Neural Information Processing Systems 24: Proceedings of the NIPS '11 conference*, pages 379–387. MIT Press, 2011.

[4] A. L. Barabasi. *Network science.* Cambridge University Press, 2016.

[5] M. Belkin, I. Matveeva, and P. Niyogi. Tikhonov regularization and semi-supervised learning on large graphs. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2004)*, pages 1000–1003, 2004.

[6] M. Bell. Alternatives to dial's logit assignment algorithm. *Transportation Research Part B: Methodological*, 29(4):287–295, 1995.

[7] D. P. Bertsekas. *Dynamic programming and optimal control, 2nd ed.* Athena Scientific, 2000.

[8] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the international Conference on Machine Learning (ICML 2001)*, pages 19–26, 2001.

[9] I. Borg and P. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 1997.

[10] M. Brand. A random walks perspective on maximizing satisfaction and profit. *Proceedings of the 2005 SIAM International Conference on Data Mining*, 2005.

[11] J. Callut, K. Francoisse, M. Saerens, and P. Dupont. Semi-supervised classification from discriminative random walks. In *Proceedings of the European conference on Machine Learning (ECML 2008)*, volume LNAI5211, pages 162–177, 2008.

[12] A. K. Chandra, P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari. The electrical resistance of a graph captures its commute and cover times. *Annual ACM Symposium on Theory of Computing*, pages 574–586, 1989.

[13] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-supervised learning*. MIT Press, 2006.

[14] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems 15: Proceedings of the NIPS '02 conference*, pages 585–592. MIT Press, 2002.

[15] P. Chebotarev. A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. *Discrete Applied Mathematics*, 159(5):295–302, 2011.

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

[17] P. Chebotarev. Studying new classes of graph metrics. In F. Nielsen and F. Barbaresco, editors, *Proceedings of the 1st International Conference on Geometric Science of Information (GSI '13)*, volume 8085 of *Lecture Notes in Computer Science*, pages 207–214. Springer, 2013.

[18] P. Chebotarev and E. Shamis. The matrix-forest theorem and measuring relations in small social groups. *Automation and Remote Control*, 58(9):1505–1514, 1997.

[19] P. Chebotarev and E. Shamis. On proximity measures for graph vertices. *Automation and Remote Control*, 59(10):1443–1459, 1998.

[20] N. Christofides. *Graph theory: An algorithmic approach*. Academic Press, 1975.

[21] F. Chung and L. Lu. *Complex Graphs and Networks*. American Mathematical Society, 2006.

[22] F. R. Chung. *Spectral graph theory*. American Mathematical Society, 1997.

[23] E. Cinlar. *Introduction to Stochastic Processes*. Prentice-Hall, 1975.

[24] J. Cook. Basic properties of the soft maximum. Unpublished manuscript available from www.johndcook.com/blog/2010/01/13/soft-maximum, 2011.

[25] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to algorithms, 3th Edition*. The MIT Press, 2009.

[26] T. Cox and M. Cox. *Multidimensional scaling, 2nd ed.* Chapman and Hall, 2001.

[27] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, Dec. 2006.

[28] R. Devooght, A. Mantrach, I. Kivimäki, H. Bersini, A. Jaimes, and M. Saerens. Random walks based modularity: Application to semi-supervised learning. In *Proceedings of the 23rd International World Wide Web Conference (WWW '14)*, pages 213–224, 2014.

[29] R. Dial. A probabilistic multipath assignment model that obviates path enumeration. *Transportation Research*, 5:83–111, 1971.

[30] P. G. Doyle and J. L. Snell. *Random Walks and Electric Networks*. The Mathematical Association of America, 1984.

[31] M. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall, 2003.

[32] E. Estrada. *The structure of complex networks*. Oxford University Press, 2012.

[33] E. Estrada and N. Hatano. Communicability in complex networks. *Physical Review E*, 77(3):036111, 2008.

[34] F. Fouss, K. Francoisse, L. Yen, A. Pirotte, and M. Saerens. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural Networks*, 31:53–72, 2012.

[35] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007.

[36] F. Fouss, L. Yen, A. Pirotte, and M. Saerens. An experimental investigation of graph kernels on a collaborative recommendation task. *Proceedings of the 6th International Conference on Data Mining (ICDM 2006)*, pages 863–868, 2006.

[37] S. García-Díez, F. Fouss, M. Shimbo, and M. Saerens. A sum-over-paths extension of edit distances accounting for all sequence alignments. *Pattern Recognition*, 44(6):1172–1182, 2011.

[38] S. García-Díez, E. Vandenbussche, and M. Saerens. A continuous-state version of discrete randomized shortest-paths. *Proceedings of the 50th IEEE International Conference on Decision and Control (IEEE CDC 2011)*, pages 6570–6577, 2011.

[39] C. Grinstead and J. L. Snell. *Introduction to probability, 2nd ed.* The Mathematical Association of America, 1997.

[40] G. Guex. Interpolating between random walks and optimal transportation routes: Flow with multiple sources and targets. *Physica A: Statistical Mechanics and its Applications*, 450:264–277, 2016.

[41] G. Guex and F. Bavaud. Flow-based dissimilarities: shortest path, commute time, max-flow and free energy. In B. Lausen, S. Krolak-Schwerdt, and M. Bohmer, editors, *Data science, learning by latent structures, and knowledge discovery*, volume 1564 of *Studies in Classification, Data Analysis, and Knowledge Organization*, pages 101–111. Springer, 2015.

[42] J. Ham, D. Lee, S. Mika, and B. Scholkopf. A kernel view of the dimensionality reduction of manifolds. *Proceedings of the 21st International Conference on Machine Learning (ICML2004)*, pages 369–376, 2004.

[43] T. Hashimoto, Y. Sun, and T. Jaakkola. From random walks to distances on unweighted graphs. In *Advances in Neural Information Processing Systems 28: Proceedings of the NIPS '15 conference*. MIT Press, 2015.

[44] M. Herbster and G. Lever. Predicting the labelling of a graph via minimum p-seminorm interpolation. *Proceedings of the 22nd Annual Conference on Learning Theory (COLT2009)*, 2009.

[45] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2088.

[46] X. Huang, Y. Ariki, and M. Jack. *Hidden Markov models for speech recognition*. Edinburgh University Press, 1990.

[47] D. Isaacson and R. Madsen. *Markov chains theory and applications*. John Wiley & Sons, 1976.

[48] T. Ito, M. Shimbo, T. Kudo, and Y. Matsumoto. Application of kernels to link analysis. *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 586–592, 2005.

[49] V. Ivashkin and P. Chebotarev. Logarithmic proximity measures outperform plain ones in graph nodes clustering. *ArXiv preprint paper, arXiv:1605.01046*, pages 1–10, 2016.

[50] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[51] T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the 20^{th} International Conference on Machine Learning (ICDM 2003)*, page 290 297, Washington DC, 2003.

[52] D. Jungnickel. *Graphs, networks, and algorithms, 3th ed.* Springer, 2008.

[53] A. Kapoor, Y. A. Qi, H. Ahn, and R. W. Picard. Hyperparameter and kernel learning for graph based semi-supervised classification. In *Advances in Neural Information Processing Systems 18: Proceedings of the NIPS '05 conference*, pages 627–634. MIT Press, 2005.

[54] H. Kappen. An introduction to stochastic control theory, path integrals and reinforcement learning. In J. Marro, P. L. Garrido, and J. J. Torres, editors, *AIP conference proceedings: ninth Granada lectures, Cooperative Behavior in Neural Systems*, volume 887 of *American Institute of Physics Conference Series*, pages 149–181, 2007.

[55] J. N. Kapur and H. K. Kesavan. *Entropy optimization principles with applications.* Academic Press, 1992.

[56] J. G. Kemeny and J. L. Snell. *Finite Markov Chains.* Springer-Verlag, 1976.

[57] I. Kivimäki, B. Lebichot, J. Saramäki, and M. Saerens. Two betweenness centrality measures based on randomized shortest paths. *Scientific Reports*, 6:srep19668, 2016.

[58] I. Kivimäki, M. Shimbo, and M. Saerens. Developments in the theory of randomized shortest paths with a comparison of graph node distances. *Physica A: Statistical Mechanics and its Applications*, 393:600–616, 2014.

[59] D. J. Klein and M. Randic. Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95, 1993.

[60] E. Kolaczyk. *Statistical analysis of network data: methods and models.* Springer, 2009.

[61] E. Kolaczyk, D. Chua, and M. Barthelemy. Group betweenness and co-betweenness: inter-related notions of coalition centrality. *Social Networks*, 31(3):190–203, 2009.

[62] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. *Proceedings of the 19th International Conference on Machine Learning (ICML 2002)*, pages 315–322, 2002.

[63] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings.* Princeton University Press, 2006.

[64] B. Lebichot, I. Kivimäki, K. Francoisse, and M. Saerens. Semi-supervised classification through the bag-of-paths group betweenness. *IEEE Transactions on Neural Networks and Learning Systems*, 25(6):1173–1186, 2014.

[65] B. Lebichot and M. Saerens. A bag-of-paths node criticality measure. *Submitted for publication*, 2017.

[66] T. G. Lewis. *Network science: theory and applications.* Wiley, 2009.

[67] M. Lichman. UCI machine learning repository, 2013.

[68] L. Lü and T. Zhou. Link prediction in complex networks: a survey. *Physica A*, 390:1150–1170, 2011.

[69] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983, 2007.

[70] C. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[71] A. Mantrach, N. van Zeebroeck, P. Francq, M. Shimbo, H. Bersini, and M. Saerens. Semi-supervised classification and betweenness computation on large, sparse, directed graphs. *Pattern Recognition*, 44(6):1212 – 1224, 2011.

[72] A. Mantrach, L. Yen, J. Callut, K. Francoise, M. Shimbo, and M. Saerens. The sum-over-paths covariance kernel: a novel covariance between nodes of a directed graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1112–1126, 2010.

[73] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press, 1979.

[74] C. D. Meyer. *Matrix analysis and applied linear algebra*. SIAM, 2000.

[75] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *Advances in Neural Information Processing Systems 18: Proceedings of the NIPS '05 conference*, pages 955–962. MIT Press, 2005.

[76] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinate of dynamical systems. *Applied and Computational Harmonic Analysis*, 21:113–127, 2006.

[77] M. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences (USA)*, 103:8577–8582, 2006.

[78] M. Newman. *Networks: an introduction*. Oxford University Press, 2010.

[79] J. R. Norris. *Markov chains*. Cambridge University Press, 1997.

[80] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. *Proceedings of the 10th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 653–658, 2004.

[81] P. Pons and M. Latapy. Computing communities in large networks using random walks. In P. Yolum, T. Gungor, F. Gurgen, and C. Ozturan, editors, *Proceedings of the 20th International Symposium on Computer and Information Sciences (ISCIS '05)*, volume 3733 of *Lecture Notes in Computer Science*, pages 284–293. Springer, 2005.

[82] P. Pons and M. Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218, 2006.

[83] A. Pucci, M. Gori, and M. Maggini. A random-walk based scoring algorithm applied to recommender engines. *Proceedings of the International Workshop on Knowledge Discovery on the Web (WebKDD 2006)*, pages 127–146, 2006.

[84] H. Qiu and E. R. Hancock. Image segmentation using commute times. *Proceedings of the 16th British Machine Vision Conference (BMVC 2005)*, pages 929–938, 2005.

[85] R. Rardin. *Optimization in operations research*. Prentice Hall, 1998.

[86] S. Ross. *Introduction to probability models, 10th ed.* Academic Press, 2010.

[87] M. Saerens, Y. Achbany, F. Fouss, and L. Yen. Randomized shortest-path problems: Two related models. *Neural Computation*, 21(8):2363–2404, 2009.

[88] M. Saerens, F. Fouss, L. Yen, and P. Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. *Proceedings of the 15th European Conference on Machine Learning (ECML 2004). Lecture Notes in Artificial Intelligence, vol. 3201, Springer-Verlag, Berlin*, pages 371–383, 2004.

[89] P. Sarkar and A. Moore. A tractable approach to finding closest truncated-commute-time neighbors in large graphs. *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[90] B. Scholkopf and A. Smola. *Learning with kernels*. The MIT Press, 2002.

[91] B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 5(10):1299–1319, 1998.

[92] R. Sedgewick. *Algorithms, 4th ed.* Addison-Wesley, 2011.

[93] T. Silva and L. Zhao. *Machine learning in complex networks*. Springer, 2016.

[94] A. J. Smola and R. Kondor. Kernels and regularization on graphs. In M. Warmuth and B. Schölkopf, editors, *Proceedings of the Conference on Learning Theory (COLT)*, pages 144–158, 2003.

[95] J. M. Steele. *Stochastic calculus and financial application*. Springer-Verlag, 2001.

[96] A. Subramanya and P. Pratim Talukdar. *Graph-based semi-supervised learning*. Morgan & Claypool Publishers, 2014.

[97] M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *Advances in Neural Information Processing Systems 14: Proceedings of the NIPS '01 conference*. MIT Press, 2001.

[98] A. Tahbaz and A. Jadbabaie. A one-parameter family of distributed consensus algorithms with boundary: from shortest paths to mean hitting times. In *Proceedings of IEEE Conference on Decision and Control*, pages 4664–4669, 2006.

[99] L. Tang and H. Liu. Relational learning via latent social dimensions. In *Proceedings of the ACM conference on Knowledge Discovery and Data Mining (KDD 2009)*, pages 817–826, 2009.

[100] L. Tang and H. Liu. Scalable learning of collective behavior based on sparse social dimensions. In *Proceedings of the ACM conference on Information and Knowledge Management (CIKM 2009)*, pages 1107–1116, 2009.

[101] L. Tang and H. Liu. Toward predicting collective behavior via social dimension extraction. *IEEE Intelligent Systems*, 25(4):19–25, 2010.

[102] H. M. Taylor and S. Karlin. *An introduction to stochastic modeling, 3th Ed.* Academic Press, 1998.

[103] M. Thelwall. *Link analysis: An information science approach*. Elsevier, 2004.

[104] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. *Proceedings of sixth IEEE International Conference on Data Mining*, pages 613–622, 2006.

[105] H. Tong, C. Faloutsos, and J.-Y. Pan. Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, 14(3):327–346, 2008.

[106] U. von Luxburg, A. Radl, and M. Hein. Getting lost in space: large sample analysis of the commute distance. In *Advances in Neural Information Processing Systems 23: Proceedings of the NIPS '10 conference*, pages 2622–2630. MIT Press, 2010.

[107] U. von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15(1):1751–1798, 2014.

[108] J. Wang, F. Wang, C. Zhang, H. Shen, and L. Quan. Linear neighborhood propagation and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1600–1615, 2009.

[109] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. Cambridge University Press, 1994.

[110] Y. Yajima and T.-F. Kuo. Efficient formulations for 1-svm and their application to recommendation tasks. *Journal of Computers*, 1(3):27–34, 2006.

[111] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens. Graph nodes clustering based on the commute-time kernel. In *Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007). Lecture notes in Computer Science*, volume LNAI4426, pages 1037–1045, 2007.

[112] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens. Graph nodes clustering with the sigmoid commute-time kernel: A comprehensive study. *Data & Knowledge Engineering*, 68(3):338–361, 2009.

[113] L. Yen, A. Mantrach, M. Shimbo, and M. Saerens. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pages 785–793, 2008.

[114] L. Yen, M. Saerens, and F. Fouss. A link analysis extension of correspondence analysis for mining relational databases. *IEEE Transactions on Knowledge and Data Engineering*, 23(4):481–495, 2011.

[115] L. Yen, D. Vanvyve, F. Wouters, F. Fouss, M. Verleysen, and M. Saerens. Clustering using a random-walk based distance measure. *Proceedings of the 13th Symposium on Artificial Neural Networks (ESANN 2005)*, pages 317–324, 2005.

[116] D. Zhang and R. Mao. Classifying networked entities with modularity kernels. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008)*, pages 113–122. ACM, 2008.

[117] D. Zhang and R. Mao. A new kernel for classification of networked entities. In *Proceedings of 6th International Workshop on Mining and Learning with Graphs*, Helsinki, Finland, 2008.

[118] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16: Proceedings of the NIPS '03 conference*, pages 237–244. MIT Press, 2003.

[119] D. Zhou, J. Huang, and B. Schölkopf. Learning from labeled and unlabeled data on a directed graph. *Proceedings of the 22nd International Conference on Machine Learning*, pages 1041–1048, 2005.

[120] D. Zhou and B. Scholkopf. Learning from labeled and unlabeled data using random walks. *Proceedings of the 26th Symposium of the German Association for Pattern Recognition (DAGM '04)*, pages 237–244, 2004.

[121] X. Zhu. Semi-supervised learning literature survey. *Manuscript available at http://pages.cs.wisc.edu/ jerryzhu/research/ssl/semireview.html*, 2008.

[122] X. Zhu and A. Goldberg. *Introduction to semi-supervised learning.* Morgan & Claypool Publishers, 2009.