A bag-of-paths graph framework with Poisson-distributed path lengths

Sylvain Courtain & Marco Saerens
Université catholique de Louvain
{sylvain.courtain,marco.saerens}@uclouvain.be

Abstract

This paper investigates a theoretical extension of the entropy-regularized least-cost problem on a graph from a bag-of-paths perspective. This extension constrains the a priori probability distribution on the length of the paths in order to follow a Poisson distribution. Therefore, this framework allows us to weigh the global impact of path lengths, depending on the structure of the graph, which proves useful in node classification and clustering problems. Accordingly, a novel distance measure between nodes of the graph can be defined from the probability of drawing an i-j path derived from the new bag-of-paths model. Experiments on supervised classification problems show that the proposed distance is competitive with other state-of-the-art distances and kernels on a graph.

1 Introduction

1.1 General introduction

This work aims at extending the *randomized shortest paths* (RSP) and *bag-of-paths* (BoP) models, which were introduced and refined in a series of papers (Bavaud & Guex, 2012; Kivimäki et al., 2014; Saerens et al., 2009; Yen et al., 2008), and were initially inspired by transportation science models (Akamatsu, 1996; Dial, 1971). Basically, the RSP model adds a relative entropy regularization term to the classical least-cost path problem between two nodes, i (source), j (target), of a graph, with the consequence that each i-j path is assigned a probability mass of following this path. Lower-cost paths are assigned a higher probability of being followed, although large-cost paths are less likely

to be chosen. This problem can also be viewed from the point of view of a maximum (relative) entropy problem with a fixed expected cost constraint (Saerens et al., 2009). The BoP model (Francoisse et al., 2017; Mantrach et al., 2010) generalizes the RSP model by extending the set of i-j paths to all possible paths in the graph (for all i-j pairs). As in Courtain & Saerens (2022), the present paper investigates a weighing of the i-j paths by the probability of choosing a given path length ℓ in the BoP framework. Here, the a priori path length distribution is assumed to follow a Poisson distribution, but other distributions could be used as well, depending on the application. This is interesting for at least two reasons. First, selecting paths having a certain length range allows us to quickly cover the entire network without relying on excessively long paths. For example, Backstrom et al. (2012), when analyzing a huge, popular social network, observed that the average length between two nodes was 4.74, corresponding to only 3.74 intermediaries or "degrees of separation". Second, the underlying intuition is that the Poisson distribution parameter could be seen as a resolution, scaling parameter monitoring the region of influence of each node (in terms of length from the starting node), which could prove useful in node clustering, community detection, or node label classification. However, the model introduced in Courtain & Saerens (2022) was derived in an ad hoc manner; therefore, it is reformulated in a more principled way in this paper, and applied in the experimental section to supervised classification problems.

1.2 Background and notation

Let us consider a weighted, strongly connected, directed graph G containing n nodes $\in \mathcal{V}$ (the set of nodes), and non-negative costs c_{ij} together with affinities a_{ij} (adjacency matrix) associated to (directed) edges.

1.2.1 The bag-of-paths model

The BoP model is based on the probability π_{ij} that a path $\wp$ drawn at random from a bag of paths starts at node i and ends at node j (Francoisse et al., 2017; Lebichot et al., 2014; Mantrach et al., 2010). This bag of paths is assumed to be a set containing all paths of G, of

arbitrary length. As usual, a path or walk $\wp$ is a sequence of transitions to adjacent nodes on G starting at a source node $s(\wp) = i$ and finishing at a target node $t(\wp) = j$. Moreover, the length of a path $\ell(\wp)$ is the number of hops required to follow that path. Each path is weighted according to its quality, that is, its total cost, defined as the sum of costs c_{kl} over all edges along the path $\wp$, $\tilde{c}(\wp)$. Costs associated with missing edges are supposed to be infinite, preventing these edges to be traversed.

Following Francoisse et al. (2017), two other notions need to be introduced to define the probability of drawing a path from the bag of paths. The first is the set of paths between a source node i and a target node j, including cycles, denoted by $\mathcal{P}_{ij} = \{\wp_{ij}\}$. The set $\mathcal{P}_{ij}$ usually contains an infinite, but countable, number of paths $\wp_{ij}$. The second is the set of all paths through the graph $\mathcal{P} = \bigcup_{i,j=1}^{n} \mathcal{P}_{ij}$.

In this context, the probability of drawing a path $\wp \in \mathcal{P}$ from the bag of paths, which is a probability distribution on the set $\mathcal{P}$, can be defined as the probability distribution $\mathrm{P}(\cdot)$ minimizing the total expected cost $\mathbb{E}[\tilde{c}(\wp)]$, favoring the *exploitation*, among all the distributions having a fixed relative entropy, or Kullback-Leibler divergence, J_0. The relative entropy is computed with respect to a reference distribution, here the natural random walk of the graph defining a Markov chain with transition probabilities $p_{kl}^{\mathrm{RW}} = a_{kl} / \sum_{l'=1}^{n} a_{kl'}$, allowing some random *exploration*.

The choice of this distribution naturally defines a probability distribution on the set of paths such that high-cost paths occur with a low probability while low-cost paths occur with a high probability (Francoisse et al., 2017). More precisely, we are seeking for path probabilities, $\mathrm{P}(\wp)$, $\wp \in \mathcal{P}$, minimizing the total expected cost subject to a constant relative entropy constraint,

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

where $J_0 > 0$ is provided a priori by the user, according to the desired degree of randomness and $\tilde{\mathrm{P}}(\wp)$ represents the probability of following

the path $\wp$ (product of probabilities along the path) when walking according to the natural random walk transition probabilities p_{kl}^{RW} gathered in transition matrix $\mathbf{P}_{\text{RW}}$, and properly normalized (Francoisse et al., 2017).

1.2.2 The path-based probability distribution

Solving the problem presented in Eq. 1 leads to a *Gibbs-Boltzmann* probability distribution (see Francoisse et al., 2017, for details),

$$\mathrm{P}(\wp) = \frac{\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp'\in\mathcal{P}}\tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \tag{2}$$

where the parameter $\theta = 1/T$ is the inverse temperature directly related to the relative entropy J_0. Thus, as expected, low-cost paths are favored due to their high probability of being sampled. The inverse temperature parameter θ allows us to monitor the balance between exploration and exploitation. Notice that, in the sequel, it will be more convenient to provide the value of the parameter θ, with $\theta > 0$, instead of the relative entropy J_0.

Finally, the bag-of-paths probability of drawing a path starting in node $s(\wp) = i$ and ending in some other node $t(\wp) = j$ can now be defined as

$$\mathrm{P}\big(s(\wp) = i, t(\wp) = j\big) = \frac{\sum_{\wp\in\mathcal{P}_{ij}}\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp'\in\mathcal{P}}\tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \tag{3}$$

with $\mathcal{P}_{ij}$ defining the set of all path starting at node i and ending at node j. As shown in Francoisse et al. (2017), this quantity can easily be computed in analytic closed-form, beginning with the introduction of a new transient (sub-stochastic) matrix $\mathbf{W}$ defined as

$$\mathbf{W} \triangleq \mathbf{P}_{\text{RW}} \circ \exp[-\theta\mathbf{C}] \tag{4}$$

where $\mathbf{C} = (c_{kl})$ is the cost matrix, $\mathbf{P}_{\text{RW}}$ is the natural transition probability matrix, $\circ$ is the elementwise (Hadamard) matrix product and the

exponential function is taken elementwise. Therefore, the element w_{kl} of $\mathbf{W}$ is $w_{kl} = p_{kl}^{\mathrm{RW}} \exp[-\theta c_{kl}]$.

Thanks to this matrix $\mathbf{W}$, it turns out (Francoisse et al., 2017) that the sum in the numerator of Eq. 3 can be rewritten as

$$\sum_{\wp \in \mathcal{P}_{ij}} \tilde{\mathrm{P}}(\wp) \exp[-\theta \tilde{c}(\wp)] = \sum_{\tau=0}^{\infty} \left[\mathbf{W}^{\tau}\right]_{ij} = \left[(\mathbf{I} - \mathbf{W})^{-1}\right]_{ij} = z_{ij} \quad (5)$$

with $\mathbf{I}$ being the identity matrix and where, by convention, zero-length paths are allowed and associated with a unit value and a zero cost. Thus, computing the power series of $\mathbf{W}$ leads to the definition of the matrix $\mathbf{Z} = (z_{kl}) \triangleq (\mathbf{I} - \mathbf{W})^{-1}$ called, by analogy to Markov chains (Kemeny et al., 1976), the *fundamental matrix*. Interestingly, it can be shown that elements z_{ij} can be interpreted as the expected number of times that a "killed" random walker with a transient matrix $\mathbf{W}$ starting from node i visits node j before stopping his walk (Francoisse et al., 2017). In the same way, the denominator of Eq. 3 can be computed by

$$\sum_{i,j=1}^{n} \sum_{\wp \in \mathcal{P}_{ij}} \tilde{\mathrm{P}}(\wp) \exp[-\theta \tilde{c}(\wp)] \triangleq \mathcal{Z} = \sum_{i,j=1}^{n} z_{ij} \quad (6)$$

with $\mathcal{Z}$ being the *partition function* of the bag-of-paths system.

1.2.3 The probability of drawing a path connecting two nodes

Finally, the *bag-of-paths probability* of drawing a path $\wp$ starting in node $s(\wp) = i$ and ending at some other node $t(\wp) = j$ (Francoisse et al., 2017), presented in Eq. 3, is

$$\pi_{ij} = \mathrm{P}\big(s(\wp) = i, t(\wp) = j\big) = \frac{z_{ij}}{\mathcal{Z}} = \frac{z_{ij}}{\sum_{i,j=1}^{n} z_{ij}} \quad (7)$$

where $\mathbf{\Pi} = (\pi_{ij})$ is the bag-of-paths probability matrix containing the probabilities for each source-target pair of nodes. Notice that this matrix verifies $\sum_{i,j=1}^{n} \pi_{ij} = 1$ and is not symmetric in general. Therefore, in the case of an undirected graph, a variant consists of computing the probability of drawing a path $i \rightsquigarrow j$ or $j \rightsquigarrow i$, i.e. regardless of the direction of the link. The result is a symmetric matrix, $\mathbf{\Pi}_{\mathrm{sym}} = \mathbf{\Pi} + \mathbf{\Pi}^{\mathrm{T}}$,

where only the upper (or lower, since the matrix is symmetric) triangular part is relevant. The interpretation of the bag-of-paths probability matrix depends on the type of path used (regular or hitting; see Kivimäki et al., 2014 and Francoisse et al., 2017 for details); in this work, we only develop the formalism for regular paths; the case of hitting paths is left for future work.

2 Bag-of-paths with Poisson-distributed path lengths

In this section, we extend the bag-of-paths model by constraining the probability of sampling a path to follow a Poisson probability distribution on its length. Moreover, this contribution also extends the content of a previous paper (Courtain & Saerens, 2022) by avoiding the independence between path likelihood and path length. First, we introduce the BoP model constraining the probability of sampling a path to be Poisson-distributed in Subsection 2.1. Then, the joint probability of drawing a path starting in node i and ending in j is derived in Subsection 2.2. Finally, a distance measure between nodes is derived from these joint probabilities in Subsection 2.3.

2.1 BoP model with Poisson-distributed path lengths

Similarly to the standard BoP model introduced in the previous section, we minimize the free energy objective function (a reformulation of our original problem in Eq. 1), while now introducing constraints on path lengths. The idea is to constrain the probability of sampling a path $\wp$ of length $\ell(\wp) = \tau$ to follow a Poisson probability distribution $f(\tau, \lambda)$ with parameter λ.[1] This additional constraint allows us to tune the expected path length at which the relevant information can be found, as a hyper-parameter. The problem in Eq. 1 can therefore be reformulated

1 Recall the form of the Poisson distribution, $f(\tau, \lambda) = \lambda^\tau \exp(-\lambda)/\tau!$ (Papoulis & Pillai, 2002). Note that other probability distributions could also be used, depending on the problem.

as

$$\left|\begin{array}{ll} \underset{\{\mathrm{P}(\wp)\}}{\text{minimize}} & \displaystyle\sum_{i,j=1}^{n}\sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\left(\mathrm{P}(\wp)\tilde{c}(\wp)+T\,\mathrm{P}(\wp)\log\left(\frac{\mathrm{P}(\wp)}{\tilde{\mathrm{P}}(\wp)}\right)\right) \\ \text{subject to} & \displaystyle\sum_{i,j=1}^{n}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\mathrm{P}(\wp)=f(\tau,\lambda) \quad \text{for each length } \tau \end{array}\right. \tag{8}$$

where, as before, $T=1/\theta$, $\tilde{c}(\wp)$ is the total cost of path $\wp$ when visiting the nodes in the sequential order and $\tilde{\mathrm{P}}(\wp)$ is the probability of the path $\wp$ according to the natural random walk. Furthermore, $\mathcal{P}_{ij}(\tau)$ is the set of paths connecting node i to node j whose length is exactly equal to τ.

Note that in this formulation of the problem, we do not explicitly constrain the probability distribution $\mathrm{P}(\wp)$ to sum to 1. Indeed, since the quantity $f(\tau,\lambda)$ is a Poisson probability mass, this implies that the probability distribution sums to 1, $\sum_{i,j=1}^{n}\sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\mathrm{P}(\wp)=\sum_{\tau=0}^{\infty}f(\tau,\lambda)=1$.

The problem presented in Eq. 8 can be solved by optimizing the following Lagrange function integrating equality constraints

$$\begin{aligned}\mathscr{L}(\mathrm{P}(\wp),\boldsymbol{\mu}) = &\sum_{i,j=1}^{n}\sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\left(\mathrm{P}(\wp)\tilde{c}(\wp)+T\,\mathrm{P}(\wp)\log\left(\frac{\mathrm{P}(\wp)}{\tilde{\mathrm{P}}(\wp)}\right)\right)\\ &+\sum_{\tau=0}^{\infty}\mu_\tau\left(f(\tau,\lambda)-\sum_{i,j=1}^{n}\sum_{\wp\in\mathcal{P}_{ij}(\tau)}\mathrm{P}(\wp)\right)\end{aligned} \tag{9}$$

over the set of path probabilities $\mathrm{P}(\wp)$ with $\wp\in\mathcal{P}=\cup_{i,j=1}^{n}\cup_{\tau=0}^{\infty}\mathcal{P}_{ij}(\tau)$ (the bag of all possible paths). Minimizing Eq. 9 can be done by setting its partial derivative to the i-j path probability $\mathrm{P}(\wp)$ of length τ to zero, which gives

$$\frac{\partial\mathscr{L}(\mathrm{P}(\wp),\boldsymbol{\mu})}{\partial\mathrm{P}(\wp)}=\tilde{c}(\wp)+T\log\left(\frac{\mathrm{P}(\wp)}{\tilde{\mathrm{P}}(\wp)}\right)+T-\mu_\tau=0 \text{ for } \wp\in\mathcal{P}_{ij}(\tau) \tag{10}$$

Isolating the logarithm and defining $\theta=1/T$ further provide

$$\mathrm{P}(\wp)=\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]\exp[\theta\mu_\tau-1] \tag{11}$$

We can now rewrite the constraint of Eq. 8, expressing that the probability of sampling a path must follow a Poisson probability distribution, from Eq. 11 as

$$f(\tau,\lambda) = \sum_{\wp\in\mathcal{P}(\tau)} \tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]\exp[\theta\mu_\tau - 1] \tag{12}$$

which means that

$$\exp[\theta\mu_\tau - 1] = \frac{f(\tau,\lambda)}{\sum_{\wp'\in\mathcal{P}(\tau)} \tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \tag{13}$$

Finally, the i-j path probabilities $\mathrm{P}(\wp)$ of length τ can be obtained from Eqs. 11 and 13 as

$$\mathrm{P}(\wp) = \frac{f(\tau,\lambda)\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp'\in\mathcal{P}(\tau)} \tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} = f(\tau,\lambda)\frac{\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\mathcal{Z}(\tau)} \tag{14}$$

where $\mathcal{Z}(\tau)$ is the partition function associated with the set of all paths with length $\ell(\wp) = \tau$, that is, $\mathcal{P}(\tau)$.

2.2 Computing the joint probability of drawing a path starting in i and ending in j

As for the standard BoP (Eq. 7), we can now define the probability of drawing a path $\wp$ starting in node $s(\wp) = i$ and ending in some other node $t(\wp) = j$, considering the set of all paths connecting i to j in exactly τ steps as $\mathcal{P}_{ij}(\tau)$. From Eq. 14, we find

$$\begin{aligned}
\pi_{ij}(\lambda) &= \mathrm{P}\big(s(\wp) = i, t(\wp) = j\big) = \sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)} \mathrm{P}(\wp) \\
&= \sum_{\tau=0}^{\infty}\sum_{\wp\in\mathcal{P}_{ij}(\tau)} \frac{f(\tau,\lambda)\tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{i',j'=1}^{n}\sum_{\wp'\in\mathcal{P}_{i'j'}(\tau)} \tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \\
&= \sum_{\tau=0}^{\infty} f(\tau,\lambda)\frac{\sum_{\wp\in\mathcal{P}_{ij}(\tau)} \tilde{\mathrm{P}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{i',j'=1}^{n}\sum_{\wp'\in\mathcal{P}_{i'j'}(\tau)} \tilde{\mathrm{P}}(\wp')\exp[-\theta\tilde{c}(\wp')]}
\end{aligned} \tag{15}$$

Furthermore, we can define, in the same way as in the standard BoP model (see Eqs. 5-6 for a given τ, or Francoisse et al., 2017), the element (i,j) of the matrix $\mathbf{Z}(\tau)$ as

$$z_{ij}(\tau) \triangleq \sum_{\wp \in \mathcal{P}_{ij}(\tau)} \underbrace{\tilde{\mathrm{P}}(\wp)\exp[-\theta \tilde{c}(\wp)]}_{w(\wp)} = \sum_{\wp \in \mathcal{P}_{ij}(\tau)} w(\wp) = [\mathbf{W}^{\tau}]_{ij} \tag{16}$$

Finally, from Eq. 16, the probability of drawing a path starting in node $s(\wp) = i$ and ending in some other node $t(\wp) = j$, presented in Eq. 15, becomes

$$\pi_{ij}(\lambda) = \mathrm{P}\big(s(\wp) = i, t(\wp) = j\big) = \sum_{\tau=0}^{\infty} f(\tau,\lambda)\frac{z_{ij}(\tau)}{z_{\bullet\bullet}(\tau)} \tag{17}$$

where $\bullet$ means summation on the corresponding index. We can immediately verify that $\sum_{i,j=1}^{n}\pi_{ij}(\lambda) = \sum_{\tau=0}^{\infty} f(\tau,\lambda) = 1$, as should be.

To obtain the probability $\pi_{ij}(\tau,\lambda)$ for an increasing length τ in matrix form, we now derive a recurrence expression to compute each term of the series presented in Eq. 17 in turn. To do so, from Eq. 16, we have $\mathbf{Z}(\tau) = \mathbf{W}^{\tau}$ so that we need to iterate the two following expressions until convergence,

$$\begin{cases} \mathbf{Z}(\tau+1) = \mathbf{Z}(\tau)\mathbf{W} \\ \mathbf{\Pi}(\tau+1,\lambda) = \mathbf{\Pi}(\tau,\lambda) + f(\tau+1,\lambda)\dfrac{\mathbf{Z}(\tau+1)}{z_{\bullet\bullet}(\tau+1)} \end{cases} \tag{18}$$

In this equation, $\mathbf{\Pi}(\tau,\lambda)$ contains the elements π_{ij} of the truncated series (Eq. 17) up to length τ. This update of $\mathbf{Z}(\tau)$ should usually converge quickly because spatial interactions in real-world networks are expected to be mainly local, which means that only low values of λ are relevant. In that situation, $f(\tau,\lambda)$ quickly drops to zero, which implies that the contributions to $\mathbf{\Pi}(\tau,\lambda)$ also tend to zero (each contribution in the series of Eq. 17 is $\leq f(\tau,\lambda)$). Furthermore, when $\tau = 0$, we initialize the matrices $\mathbf{Z}(\tau)$ and $\mathbf{\Pi}(\tau,\lambda)$ by

$$\begin{cases} \mathbf{Z}(0) = \mathbf{I} \\ \mathbf{\Pi}(0,\lambda) = f(0,\lambda)\dfrac{\mathbf{I}}{n} \end{cases} \tag{19}$$

with n being the number of nodes and $\mathbf{I}$ the identity matrix. This means that, for zero-length paths, the source node and the target node must be the same, and thus, $\pi_{ij}(0, \lambda) = f(0, \lambda)\delta_{ij}/n$, where δ_{ij} is the Kronecker delta.

The time complexity of the matrix $\mathbf{\Pi}(\tau, \lambda)$ computation is dominated by the matrix product performed at each iteration. Therefore, it is of order $k \cdot \mathcal{O}(n^3)$ where n is the number of nodes and k is the number of required iterations. However, complexity could be lower when working with sparse matrices.

2.3 A derived distance measure between nodes

We now derive a distance measure between the nodes following the same procedure as in Courtain & Saerens (2022) and Francoisse et al. (2017). More specifically, we take minus the (elementwise) logarithm of the probability matrix $\mathbf{\Pi}(\tau, \lambda)$ obtained after convergence of Eq. 18. The resulting distance matrix between nodes, called the *directed Poisson surprisal distance*, is defined as $\mathbf{\Delta}^{\text{DPSURP}} = -\log \mathbf{\Pi}(\tau, \lambda)$ with the log being the natural elementwise logarithm. It computes the "surprisal" of observing a path starting in i and ending in j, and is an extension of the surprisal distance introduced in Francoisse et al. (2017). The *Poisson surprisal distance* (PSurp) measure is the symmetrized quantity, $\mathbf{\Delta}^{\text{PSURP}} = \frac{1}{2}\left(\mathbf{\Delta}^{\text{DPSURP}} + (\mathbf{\Delta}^{\text{DPSURP}})^{\text{T}}\right)$, where the diagonal elements are then set to zero by subtracting $\mathbf{Diag}(\mathbf{\Delta}^{\text{PSURP}})$ to $\mathbf{\Delta}^{\text{PSURP}}$. It measures both proximity (low cost) and reachability (high connectivity) of the nodes of G; in other words, it quantifies the ease of accessibility between pairs of nodes.

3 Experiments

In this section, we compare the performance of the Poisson surprisal distance introduced in the previous section with other state-of-the-art methods regarding classification accuracy on a graph-based kernel semi-supervised classification task. It is important to emphasize that our description follows the framework established in Courtain (2022) and Courtain & Saerens (2022), and that the compared methods, experimental setup, and datasets analyzed are similar to those used in that study.

Therefore, we will only focus on the new points in our description and refer to those papers for more details.

3.1 Investigated state-of-the-art methods

As part of our experimental design, we selected three dissimilarity measures and seven kernel matrices as baseline methods to evaluate the performance of the introduced distance. These methods achieved the highest performance in their respective categories in the semi-supervised classification experiment described in Courtain (2022). Furthermore, the majority of these methods have already demonstrated strong performance in prior semi-supervised tasks (Courtain et al., 2023; Courtain & Saerens, 2022; Francoisse et al., 2017; Ivashkin & Chebotarev, 2022; Leleux et al., 2021), as well as in unsupervised tasks (Courtain et al., 2021; Ivashkin & Chebotarev, 2017; Sommer et al., 2017; Yen et al., 2009).

A summary of all the investigated methods and their acronyms is provided in table 1; notice that our proposed method is called PSurp. We refer the interested reader to subsection 7.1 and table 7.1 in Courtain, 2022 for an in-depth description of each method and the parameter values used. For our proposed method, we employed the same parameters as PWSurp (introduced in Courtain & Saerens, 2022), specifically $\theta = \{10^{-6}, 10^{-5}, \ldots, 10\}$ and $\lambda = \{1, 2, 3, 5, 10\}$. Furthermore, for PWSurp, we employed both uniform priors (PWSurpUni) and L_1-normalized degree priors (PWSurpDegree) to enable direct comparison with PSurp, even though previous studies have shown that L_1-normalized priors yield superior performance (Courtain, 2022; Courtain & Saerens, 2022).

3.2 Experimental design

To evaluate node classification performance across the 14 network datasets, we transform all dissimilarity and similarity measures into kernel matrices, $\mathbf{K}$, by removing negative eigenvalues and feeding these matrices into a kernel SVM[2] with various margin parameter values $c = \{10^{-2}, 10^{-1}, 1, 10, 100\}$. As noted earlier, the same dataset

2 The LIBSVM library (Chang & Lin, 2011) was used with options '-s 0' and '-t 4'.

Method	Acronym
Poisson surprisal distance (this paper)	PSurp
Poisson-weighted surprisal distance with priors (by Uniform and Degree distribution) (Courtain & Saerens, 2022)	PWSurp
Margin-constrained bag-of-hitting-paths surprisal distance (Guex et al., 2019)	cBoPH
Correlation kernel based on the number of occurrences of nodes on regular paths (Guex et al., 2021)	nCor
Correlation kernel based on the number of occurrences of nodes on hitting paths (Guex et al., 2021)	nCorH
Logarithmic forest distance (Chebotarev, 2011)	LF
Modified regularized Laplacian kernel (Ito et al., 2005)	MRL
Sigmoid commute time similarity (Yen et al., 2007, 2009)	SCT
Sigmoid corrected commute time similarity (based on von Luxburg et al., 2010, Yen et al., 2007, 2009)	SCCT
Logarithmic communicability similarity (Ivashkin & Chebotarev, 2017)	LogCom
Random walk with restart similarity (Tong et al., 2006)	RWWR

TABLE 1 – The different methods for computing similarities and dissimilarities between nodes investigated in our experiments, with their acronym.

collection as in Courtain (2022) is employed; detailed descriptions can be found in table 7.3 of that work. This collection comprises nine subsets from the 20 Newsgroup datasets (Lichman, 2013; Yen et al., 2009), four WebKB datasets (Macskassy & Provost, 2007), and the IMDB dataset (Macskassy & Provost, 2007). All networks are undirected, represented by an adjacency matrix $\mathbf{A}$, with transition costs c_{ij} defined as the inverse of affinity, $1/a_{ij}$, akin to electrical networks (Francoisse et al., 2017). Each node is labeled with a class for classification purposes based on the structure of the network.

For dissimilarity measures, we examine three transformations to a kernel matrix: classical multidimensional scaling (MDS)[3] (Borg & Groenen, 1997), Gaussian transformation (Gauss) (Schölkopf & Smola, 2002), and centered Gaussian transformation (GaussCenter).[4] To maintain clarity, we present only the best kernel transformation results for each method, as determined by Nemenyi tests (Demšar, 2006) (see Tab. 2 for details).

3 Kernels generated through this transformation are inherently centered.

4 In this transformation, the kernel $\mathbf{K}$ is centered as $\mathbf{K} \leftarrow \mathbf{HKH}$, where $\mathbf{H} = \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}}/n$ is the centering matrix, $\mathbf{e}$ is a vector of ones, $\mathbf{I}$ is the identity matrix, and n is the number of nodes.

To minimize variability in the results, we conduct 10 repetitions of a 5×5 nested cross-validation procedure, with different labeled and unlabeled node splits in each run. In the external 5-fold cross-validation, 20% of the node labels are used for training, while the remaining 80% are hidden for testing. During each internal 5-fold cross-validation, parameters are tuned on the external training fold using 80% of the labeled nodes. External and internal folds are kept consistent across all methods in a given run to maintain comparability. The final results, presented in table 2, are the average of 50 accuracy scores from the external cross-validation folds.

3.3 Results and discussion

The results of the semi-supervised classification experiments are presented in table 2. To enhance clarity, the highest accuracy for each dataset is indicated in bold. As shown in table 2, LF outperforms all others on the WebKB datasets, while the nCor method achieves the best performance on the IMDB dataset. The results for the Newsgroups datasets are more varied: PWSurpDegree and SCCT each deliver the highest accuracy on three of these datasets, whereas PSurp, PWSurpUni, and cBoPH each attain the highest accuracy on one dataset. Overall, these results suggest that the best-performing method varies depending on the dataset.

To further investigate the results, we first conducted a nonparametric Friedman-Nemenyi statistical test, followed by multiple Wilcoxon signed-rank tests with a 95% confidence level ($\alpha = 0.05$) (Demšar, 2006) based on the average accuracy computed on the 14 datasets. The Friedman test yielded a p-value of 2.5×10^{-12}, which is below the α threshold, indicating that at least one method's performance is significantly different from the others. Given the positive result of the Friedman test, we proceeded with the Nemenyi test, with the results shown in Figure 1. To refine our analysis and provide deeper insights into the relative performance of the methods, we also conducted multiple Wilcoxon signed-rank tests, the results of which are summarized in table 3.

Dataset	PSurp	PWSurpUni	PWSurpDeg	CBoPH	nCor	nCorH
WebKB-Texas	77.37±3.94	78.44±3.14	78.55±3.13	76.68±3.26	67.36±2.47	68.01±2.10
WebKB-Washington	70.41±1.99	70.42±2.02	71.85±2.04	70.12±2.19	65.84±0.97	65.71±1.23
WebKB-Wisconsin	77.31±2.19	77.08±2.46	77.95±2.14	76.42±2.42	73.74±1.85	73.70±1.78
WebKB-Cornell	59.47±2.91	59.56±2.61	59.61±2.59	58.79±3.38	52.89±3.63	53.58±3.51
IMDB	77.57±1.36	77.52±1.47	77.95±1.33	79.24±1.48	**79.63±1.16**	79.60±1.10
Newsgroup-2cl-1	96.21±1.13	96.36±1.40	96.13±1.12	95.63±0.88	95.51±1.09	95.32±1.18
Newsgroup-2cl-2	**92.76±1.64**	92.50±1.82	92.41±1.78	92.59±1.60	92.47±1.57	92.56±1.72
Newsgroup-2cl-3	96.53±0.95	96.48±1.02	96.44±0.99	**96.62±0.74**	95.93±1.28	96.05±1.23
Newsgroup-3cl-1	93.37±1.22	93.39±1.12	93.56±1.06	93.33±1.00	92.98±1.43	92.84±1.26
Newsgroup-3cl-2	93.34±1.18	93.32±1.11	93.23±1.05	93.23±1.00	92.39±1.06	92.28±1.14
Newsgroup-3cl-3	92.74±1.23	92.78±1.35	**93.28±1.32**	93.26±1.02	92.43±1.35	92.45±1.27
Newsgroup-5cl-1	89.15±0.99	**89.37±1.02**	89.25±1.07	88.94±0.89	87.87±1.49	87.93±1.54
Newsgroup-5cl-2	84.12±1.50	84.38±1.42	**84.46±1.52**	83.94±1.32	82.78±1.42	82.76±1.42
Newsgroup-5cl-3	83.80±1.83	83.67±1.91	**84.15±1.51**	83.54±1.32	82.36±1.28	82.49±1.30

Dataset	LF	MRL	SCT	SCCT	LogCom	RWWR
WebKB-Texas	**79.46±2.72**	49.30±1.95	76.54±3.49	76.80±3.01	79.04±2.47	51.08±5.68
WebKB-Washington	**71.87±1.99**	64.95±1.39	69.49±2.49	68.66±2.18	71.35±2.41	65.36±1.36
WebKB-Wisconsin	**79.48±1.95**	50.12±0.91	77.70±2.29	77.64±2.03	78.44±2.26	63.11±4.35
WebKB-Cornell	**59.96±2.65**	41.91±0.09	58.92±2.89	59.13±2.94	58.85±2.88	40.22±5.65
IMDB	79.19±1.36	77.82±2.42	78.03±1.65	77.41±1.43	77.29±1.64	79.05±1.10
Newsgroup-2cl-1	94.95±1.72	94.46±1.64	95.84±1.03	**97.07±0.67**	95.80±1.31	94.42±1.76
Newsgroup-2cl-2	92.27±1.54	91.70±1.70	92.22±1.41	92.50±1.29	92.32±1.55	91.98±1.81
Newsgroup-2cl-3	95.53±1.56	95.02±1.26	96.00±1.00	96.06±0.75	95.27±1.59	95.65±1.19
Newsgroup-3cl-1	92.70±1.13	91.91±1.51	93.43±1.02	**93.94±0.63**	93.28±1.11	92.74±1.19
Newsgroup-3cl-2	92.35±1.33	91.52±1.50	92.25±1.15	**93.36±0.85**	92.36±1.00	92.29±1.36
Newsgroup-3cl-3	92.36±1.43	91.26±1.56	92.64±1.12	93.11±0.78	91.61±1.19	91.26±1.73
Newsgroup-5cl-1	87.95±1.17	86.26±1.67	86.96±1.15	88.00±1.03	87.77±1.46	87.48±1.26
Newsgroup-5cl-2	82.69±1.84	80.94±2.03	81.86±1.59	82.96±0.85	83.16±1.51	82.77±1.43
Newsgroup-5cl-3	82.50±1.72	80.94±1.71	82.17±1.77	83.56±1.18	82.73±1.79	82.18±1.52

TABLE 2 – Classification accuracy in percentage terms (mean ± standard deviation) for the various classification methods across different datasets. For each dataset and method, the final accuracy is computed following the experimental design outlined in Subsection 3.2. The best-performing method for each dataset is highlighted in bold.

The Nemenyi test indicates that MRL and RWWR perform significantly worse than PSurp, PWSurpUni, PWSurpDegree, cBoPH, and SCCT. Additionally, MRL is also outperformed by LF and LogCom. The test further shows that PWSurpDegree performs significantly better than SCT, nCor, and nCorH.

In addition to confirming the findings of the Nemenyi test, the Wilcoxon tests (see Tab. 3) reveal that PWSurpDegree outperforms all the evaluated methods except for LF and PWSurpUni. The tests further indicate that PSurp, PWSurpUni, cBoPH, and SCCT outperform nCor, nCorH, and SCT. Moreover, they also show that MRL performs worse than all other methods in the analysis. Finally, RWWR is outperformed by all methods, except for SCT and MRL.

Overall, the experiments demonstrated that the newly introduced dis-

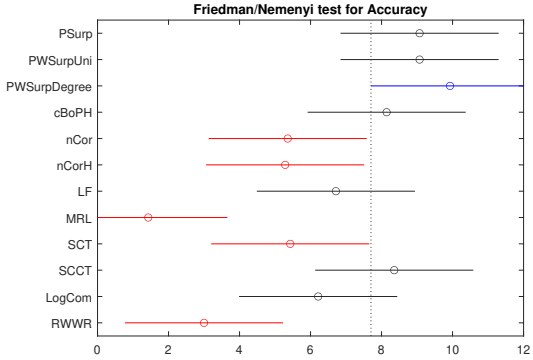

FIGURE 1 – Mean ranks and 95% Nemenyi confidence intervals for the 12 methods evaluated across 14 datasets. Significant differences between methods are determined by non-overlapping confidence intervals. The x-axis represents the average rank of each method, where a higher rank indicates better performance. The top-performing method (PWSurpDegree) and the lowest-ranked methods (nCor, nCorH, MRL, SCT, and RWWR) are highlighted.

Kernel → **Kernel ↓**	PSurp	PWSurpUni	PWSurpDegree	cBoPH	nCor	nCorH	LF	MRL	SCT	SCCT	LogCom	RWWR
PSurp	1.0000	0.7148	**0.0166**	0.0906	**0.0052**	**0.0052**	1.0000	**0.0002**	**0.0052**	0.2676	0.1726	**0.0023**
PWSurpUni	0.7148	1.0000	0.1041	0.1531	**0.0052**	**0.0067**	0.7148	**0.0002**	**0.0134**	0.2958	0.0906	**0.0023**
PWSurpDegree	**0.0166**	0.1041	1.0000	**0.0295**	**0.0040**	**0.0040**	0.2166	**0.0001**	**0.0002**	**0.0353**	**0.0040**	**0.0012**
cBoPH	0.0906	0.1531	**0.0295**	1.0000	**0.0009**	**0.0006**	0.8552	**0.0001**	**0.0295**	0.6698	0.5016	**0.0001**
nCor	**0.0052**	**0.0052**	**0.0040**	**0.0009**	1.0000	0.8077	0.7609	**0.0001**	0.3575	**0.0052**	0.2412	**0.0001**
nCorH	**0.0052**	**0.0067**	**0.0040**	**0.0006**	0.8077	1.0000	0.8077	**0.0001**	0.4263	**0.0085**	0.2166	**0.0006**
LF	1.0000	0.7148	0.2166	0.8552	0.7609	0.8077	1.0000	**0.0001**	0.0580	0.9515	0.2676	**0.0031**
MRL	**0.0002**	**0.0002**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	1.0000	**0.0001**	**0.0002**	**0.0006**	**0.0107**
SCT	**0.0052**	**0.0134**	**0.0002**	**0.0295**	0.3575	0.4263	0.0580	**0.0001**	1.0000	**0.0353**	0.2958	0.0676
SCCT	0.2676	0.2958	**0.0353**	0.6698	**0.0052**	**0.0085**	0.9515	**0.0002**	**0.0353**	1.0000	0.3910	**0.0031**
LogCom	0.1726	0.0906	**0.0040**	0.5016	0.2412	0.2166	0.2676	**0.0006**	0.2958	0.3910	1.0000	**0.0166**
RWWR	**0.0023**	**0.0023**	**0.0012**	**0.0001**	**0.0001**	**0.0006**	**0.0031**	**0.0107**	0.0676	**0.0031**	**0.0166**	1.0000

TABLE 3 – The p-values are reported from pairwise Wilcoxon signed-rank tests applied to the results presented in table 2. The p-values below the threshold of 0.05 are highlighted to indicate statistical significance.

tance achieved strong performance in our experimental setup. It proved to be competitive with cBoPH and SCCT, which have consistently shown good results across a wide range of experiments in previous semi-supervised (Courtain et al., 2023; Courtain & Saerens, 2022; Francoisse et al., 2017; Ivashkin & Chebotarev, 2022; Leleux et al., 2021) and unsupervised tasks (Courtain et al., 2021; Ivashkin & Chebotarev, 2017; Sommer et al., 2017; Yen et al., 2009).

Additionally, in line with the findings of Courtain & Saerens (2022), PWSurpDegree once again emerges as the most effective method. How-

ever, unlike PWSurpUni and PSurp, this approach incorporates priors, which could explain the observed differences in performance. Notably, experiments have demonstrated that both methods (PSurp and PWSurp) show very similar performance levels when using uniform priors. This observation was previously emphasized in an exploratory analysis, which revealed that these two distances have a correlation exceeding 90% (see Subsection 6.3.4 in Courtain, 2022).

4 Conclusions and future work

This paper investigated a mechanism constraining the probability of drawing a path from a bag of paths to follow a predefined discrete probability distribution on their length (illustrated here with a Poisson distribution). Consequently, the marginal probability of selecting a path of a given length follows this specified distribution. The introduction of this path-length distribution extends the basic BoP framework by allowing for a more precise tuning of the model, according to the application under study.

More precisely, an algorithm computing the probability of drawing a path connecting a given source and a given target node is developed. Then, taking minus the logarithm of this probability provides a dissimilarity measure between the two nodes, in terms of accessibility in the network, called the Poisson surprisal distance. This dissimilarity measure is then investigated in an experimental comparison with other state-of-the-art algorithms. The introduced measure was shown to provide competitive results.

Future work will consider other path length probability distributions (instead of Poisson), but also the introduction of a priori probabilities at source and target nodes, which showed superior results in previous work (Courtain & Saerens, 2022). In addition, it would also be interesting to compute the expected cost between two nodes within the same formalism, which would also provide a new dissimilarity measure between nodes.

Acknowledgments

We thank Professor François Bavaud of the University of Lausanne (Switzerland) and an anonymous reviewer for their comments and suggestions.

References

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

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, page 33–42, New York, NY, USA. Association for Computing Machinery.

Bavaud, F. & Guex, G. (2012). Interpolating between random walks and shortest paths: A path functional approach. In Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., & Guéret, C. (Eds.), *Proceedings of the 4th International Conference on Social Informatics (SocInfo '12)*, volume 7710 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

Borg, I. & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. Springer, New York, NY, USA.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

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

Courtain, S. (2022). *Essays on network data analysis through the bag-of-paths framework*. PhD thesis, Université catholique de Louvain, Belgium.

Courtain, S., Guex, G., Kivimäki, I., & Saerens, M. (2023). Relative entropy-regularized optimal transport on a graph: a new algorithm and an experimental comparison. *International Journal of Machine Learning and Cybernetics*, 14(4):1365–1390.

Courtain, S., Leleux, P., Kivimäki, I., Guex, G., & Saerens, M. (2021). Randomized shortest paths with net flows and capacity constraints. *Information Sciences*, 556:341–360.

Courtain, S. & Saerens, M. (2022). A simple extension of the bag-of-paths model weighting path lengths by a Poisson distribution. In *Complex Networks & Their Applications X*, pages 220–233, Cham. Springer International Publishing.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

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

Francoisse, K., Kivimäki, I., Mantrach, A., Rossi, F., & Saerens, M. (2017). A bag-of-paths framework for network data analysis. *Neural Networks*, 90:90–111.

Guex, G., Courtain, S., & Saerens, M. (2021). Covariance and correlation kernels on a graph in the generalized bag-of-paths formalism. *Journal of Complex Networks*, 8(6).

Guex, G., Kivimäki, I., & Saerens, M. (2019). Randomized optimal transport on a graph: framework and new distance measures. *Network Science*, 7(1):88–122.

Ito, T., Shimbo, M., Kudo, T., & Matsumoto, Y. (2005). Application of kernels to link analysis. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '05)*, pages 586–592.

Ivashkin, V. & Chebotarev, P. (2017). Do logarithmic proximity measures outperform plain ones in graph clustering? In Kalyagin, V. A., Nikolaev, A. I., Pardalos, P. M., & Prokopyev, O. A. (Eds.), *Models, algorithms, and technologies for network analysis*, pages 87–105, Cham. Springer International Publishing.

Ivashkin, V. & Chebotarev, P. (2022). Dissecting graph measure performance for node clustering in LFR parameter space. In *Proceedings of the 10th International Conference on Complex Networks and their Applications (CNA '21)*, pages 328–341. Springer.

Kemeny, J. G., Snell, J. L., & Knapp, A. (1976). *Denumerable Markov chains*. Springer, New York, NY, USA.

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

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

Leleux, P., Courtain, S., Guex, G., & Saerens, M. (2021). Sparse randomized shortest paths routing with Tsallis divergence regularization. *Data Mining and Knowledge Discovery*, 35:986–1031.

Lichman, M. (2013). UCI machine learning repository.

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

Mantrach, A., Yen, L., Callut, J., Francoise, K., Shimbo, M., & Saerens, M. (2010). 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.

Papoulis, A. & Pillai, S. U. (2002). *Probability, random variables and stochastic processes*. McGraw-Hill, New York, NY, USA, 4th edition.

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

Schölkopf, B. & Smola, A. (2002). *Learning with kernels*. MIT Press, Cambridge, MA, USA.

Sommer, F., Fouss, F., & Saerens, M. (2017). Modularity-driven kernel k-means for community detection. In Lintas, A., Rovetta, S., Verschure, P. F., & Villa, A. E. (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2017*, pages 423–433, Cham. Springer International Publishing.

Tong, H., Faloutsos, C., & Pan, J.-Y. (2006). Fast random walk with restart and its applications. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM '06)*, pages 613–622.

von Luxburg, U., Radl, A., & Hein, M. (2010). Getting lost in space: Large sample analysis of the commute distance. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS '10)*, pages 2622–2630. MIT Press.

Yen, L., Fouss, F., Decaestecker, C., Francq, P., & Saerens, M. (2007). Graph nodes clustering based on the commute-time kernel. In Zhou, Z.-H., Li, H., & Yang, Q. (Eds.), *Advances in Knowledge Discovery and Data Mining*, pages 1037–1045, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Yen, L., Saerens, M., Mantrach, A., & Shimbo, M. (2008). A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 785–793, New York, NY, USA. Association for Computing Machinery.