

# Sparse Randomized Shortest Paths Routing with Tsallis Divergence Regularization

*Draft manuscript subject to changes*

Pierre Leleux, Sylvain Courtain, Guillaume Guex & Marco Saerens

## Abstract

This work elaborates on the important problem of (1) designing optimal randomized routing policies for reaching a target node  $t$  from a source node  $s$  on a weighted directed graph  $G$  and (2) defining distance measures between nodes interpolating between the least cost (based on optimal movements) and the commute-cost (based on a random walk on  $G$ ), depending on a temperature parameter  $T$ . To this end, the randomized shortest path formalism (RSP, [2, 99, 124]) is rephrased in terms of Tsallis divergence regularization, instead of Kullback-Leibler divergence. The main consequence of this change is that the resulting routing policy (local transition probabilities) becomes sparser when  $T$  decreases, therefore inducing a sparse random walk on  $G$  converging to the least-cost directed acyclic graph when  $T \rightarrow 0^+$ . Experimental comparisons on node clustering and semi-supervised classification tasks show that the derived dissimilarity measures based on expected routing costs provide state-of-the-art results. The sparse RSP is therefore a promising model of movements on a graph, balancing sparse exploitation and exploration in an optimal way.

## 1 Introduction

### 1.1 General introduction

Link analysis and network science are currently used in a large number of different fields for analysing network data, like social networks, networks of transactions, protein networks, road networks, etc (see, e.g., [6, 12, 21, 31, 34, 65, 72, 88, 101, 110, 120]). In this context, many important problems arise, such as the definition of meaningful distance measures between nodes or models of movement / communication in the network (routing policies). These quantities usually take the structure of the whole network into account.

One such model is the **randomized shortest paths (RSP)**; see the related work), first developed in transportation sciences [2] and then further extended in the context of network data analysis [36, 63, 99, 124]. Assume an agent walking on the graph who wants to reach a target node  $t$  from some source node  $s$  by following some path or walk connecting these two nodes. The RSP uses a statistical physics formalism by putting a Gibbs-Boltzmann probability distribution on the set of paths<sup>1</sup> from  $s$  to  $t$ , depending on a temperature parameter  $T$ . The agent then chooses a path

---

<sup>1</sup>Considered as independent.

to follow according to this Gibbs-Boltzmann distribution – a *global policy* or *routing strategy* in terms of paths. When  $T$  is low, low-cost paths are favored while when  $T$  is large, paths are chosen according to their likelihood in a completely random walk on the graph  $G$ . It has been shown that this model of movement defines, at the local node level, a *biased random walk* on  $G$  attracting the walker to the target node. This biased random walk defines the *local policy* captured by the transition probabilities matrix of an absorbing Markov chain. The model can be derived by minimizing expected path cost regularized by Kullback-Leibler (**KL**) divergence based on path probabilities, and thus provides an optimal policy in that sense (see the above references for details).

Inspired by [3, 7, 99], this work first shows how the paths-based minimization problem can be rephrased in terms of a local objective function, that is, the cost function and the regularization term are expressed in terms of transition probabilities instead of path probabilities. This reformulation allows to replace the KL divergence by the Tsallis divergence ([115, 116], also called Havrda-Charvat divergence [50, 60]), with the consequence that the obtained policy becomes *sparse* for low temperatures (as, e.g., in [73, 74]; see the related work). In these conditions, the resulting biased random walk is only defined on a subset of edges (a subgraph of  $G$ ) and when  $T$  decreases toward zero, the most costly paths are gradually removed until only shortest paths remain active (see the illustration on Figure 2). More generally, the model of movement interpolates between a pure random walk (large  $T$ , exploration) and the least cost paths (low  $T$ , exploitation). The proposed algorithm iteratively computing the policy is inspired by [57, 85, 99] and solves the primal problem, providing transition probabilities, and then the dual problem for computing the Lagrange parameters which correspond to the minimized free energy, until convergence.

In addition, two new *dissimilarity measures* between nodes are derived from this framework, the Tsallis RSP and **free energy (FE)** dissimilarities<sup>2</sup>. These are the counterparts of the same dissimilarities defined in the standard RSP framework based on the KL divergence: the RSP is the expected cost from  $s$  to  $t$  when following the optimal randomized policy and the FE is the minimized free energy objective function at the optimal routing policy [63]. Both capture the notion of *relative accessibility* (proximity and amount of inter-connectivity [20]) between nodes and interpolate (up to a scaling factor) between the least-cost and the commute-cost distances on undirected graphs.

## 1.2 Related work

Traditional network measures are usually derived from two different paradigms about movement, or communication, taking place in the network [46]: optimal communication based on least cost paths, and random communication based on a random walk on the graph. For instance, the shortest path distance and the standard betweenness centrality [38] are defined from least cost paths whereas resistance distance and random walk centrality [13, 87] rely on random walks [29]. But, in practice, movements over a network hardly ever occur either perfectly optimally or perfectly randomly: the agent usually has some, although incomplete, knowledge of his environment. As already mentioned, the RSP framework [36, 63, 99, 124] relaxes these assumptions by interpolating between least cost paths and a pure random walk on the graph, depending on a temperature parameter. In this context, the RSP has been used recently, e.g., for modelling the behavior of animals during migration [91]. Initially, the model was inspired by transportation models developed in transportation sciences [2], and is also closely related to the work of [113, 114] in reinforcement learning and [26] defining random walks on a graph with minimal free energy or maximal entropy.

---

<sup>2</sup>We conjecture that the Tsallis FE is a distance measure as it verified the properties of a distance measure in all our applications.

Besides these previous works, other interesting models of movement interpolating between random and optimal behaviors have been proposed in the recent literature. Some of these models are based on electrical networks and extensions [4, 53, 89, 118], others on combinatorial analysis arguments [16, 17, 18], on L1-L2 regularization [73, 74], as well as on network flow models with Kullback-Leibler divergence regularization [7, 44]. These last propositions are equivalent to the RSP, although derived from another perspective. Many of these models lead to distance measures between nodes avoiding the so-called “lost in space” effect [117, 118] by interpolating between the least cost distance and the resistance distance (equivalent to the commute-time distance and the commute-cost distance, up to a scaling factor [33]). For a more thorough discussion of these related works, see, e.g., [34, 36].

Among these references, the most closely related works are [73, 74]. The authors present a sparse routing strategy based on edge flow optimization regularized by mixed L1/L2 norms [49]. More precisely, they use the edge flow formalism [1, 28] and minimize the weighted sum of (non-negative) net flows and squared net flows subject to flow conservation constraints. Although these papers address problems similar to our work and are of high interest, there are significant differences. First, they consider undirected networks while we address directed ones. They also consider net flows (like in electrical networks) while the present work considers raw flows. Moreover, the regularization technique is different: in [73, 74], a mixed L1/L2 norm is minimized whereas we use a Tsallis divergence regularization term<sup>3</sup>. Therefore, the derived algorithms are different. In addition, our work extends the scope of the RSP framework by using the Tsallis divergence instead of the KL divergence, whereas the [73, 74] work is based on network flow theory. Furthermore, one of our main goals is to derive dissimilarity measures between nodes, which is not addressed in [73, 74].

Concerning Tsallis entropy regularization, it is known for years that when minimizing the regularized expected cost together with sum-to-one and non-negativity constraints, the resulting solution becomes sparse. This is not surprising because the Tsallis entropy, in its basic form, is the L2 squared norm of a non-negative vector whose entries sum to one (a probability mass). It is therefore closely related to L1/L2 regularization which is known to provide sparse solutions [14, 48, 49]. For instance, it was shown in [57, 85] that this technique provides sparse discrete cluster membership probabilities in fuzzy clustering. In this context, a discussion of the use of various regularization terms in fuzzy clustering is provided in [82, 84]; however, [82] considers a different objective function which does not lead to sparse solutions. Another interesting application is the sparse probabilistic latent semantic analysis introduced in [51]. The author proposes an expectation-maximization algorithm based on Tsallis divergence, instead of KL divergence. In another paper [52], the same authors study maximum Tsallis entropy estimation and propose an algorithm solving the problem. Interestingly, the paper also discusses the different ways of defining maximum Tsallis entropy (see [52], subsection 1.2), with only one of these ways being relevant in our applications. Note also that in [58], Kanzawa considers Tsallis divergence regularization, but again uses a different objective function that does not lead to a sparse solution.

Also closely related to our work, the recent interesting paper [71] investigates sparse policies for discounted Markov decision processes (**MDP**), with application to reinforcement learning (see also [41]). As initiated in [113, 114] (and related to the alternative view of the RSP presented in Subsection 2.3), the authors add an entropy term to the cost associated to each action choice in the MDP. Instead of using the KL divergence as in [113], they choose the Tsallis entropy with  $r = 2$  and a uniform reference probability distribution (see Equation (8) below). As in [57, 85], they show that the resulting policy enforces sparsity. This problem is in fact closely connected to

---

<sup>3</sup>See the discussion following Equation (10) for details about the differences between the two objective functions.

the recently introduced sparse maximum procedure [67, 81] and the probability simplex projection, e.g. [8, 22, 30, 119] who proposed procedures similar to [57, 85] for solving this problem. Our work therefore solves a problem similar to [71] because a Markov chain is a particular case of a MDP. However, our results are complementary, as the optimal policy is derived from a different perspective (the RSP). Moreover, they extend to general Tsallis divergences of the form provided by Equation (8), with  $r > 1$  (generalizing Tsallis entropy), and thus also integrating a reference distribution. Therefore, these new results could easily be applied to the MDP problem, which will be considered in further work. Note also that we are working with absorbing Markov chains while [71] investigates a discounted process, but this is mainly a detail. Finally, as already stressed, our objective, in addition to derive optimal routing policies, is to introduce new dissimilarity measures interpolating between the least-cost and the commute-cost distances between nodes of a graph.

Other recent works related to the induced sparsity of Tsallis regularization are [81] for obtaining sparse outputs of a classifier in the context of multi-label classification [70], extending [71] and proposing a “maximum causal Tsallis entropy” framework (see [71] for details). Finally, in [86], the authors unify the two main approaches to optimal transport, namely Monge-Kantorovitch and Sinkhorn-Cuturi, into what they define the Tsallis-regularized optimal transport. As the name indicates, the objective function is regularized by Tsallis  $r$ -entropy and they then analyse the optimization problem by using the  $r$ -exponential formalism [116]. Then, a different approach from the one introduced in this paper is used in order to solve the problem, namely a sophisticated gradient-based algorithm.

### 1.3 Contributions and contents of the paper

In short, the main contributions of this paper are

- ▶ The development and the investigation of a new randomized shortest paths model of movement on a graph, considering Tsallis divergence instead of the KL divergence. This model shows some interesting properties: it provides sparse routing policies and it interpolates between least-cost and random-walk routing, depending on the temperature parameter  $T$ .
- ▶ A procedure for computing the optimal policy (sparse when  $T$  is small), minimizing expected cost plus Tsallis divergence, is derived. It therefore extends previous results by minimizing the Tsallis divergence instead of the Tsallis entropy.
- ▶ Two new dissimilarity measures between nodes, interpolating<sup>4</sup> between the least cost and the commute-cost distances based on a model of sparse movements on the network are introduced.
- ▶ It provides an experimental comparison between the two new dissimilarities based on Tsallis divergence regularization and other baseline dissimilarities on nodes clustering and semi-supervised classification tasks.

The content of the paper is as follows. Section 2 provides a brief introduction to the RSP framework and introduces an alternative view on the RSP which can easily be generalized by considering other entropic regularizations. Then, in Section 3, the sparse RSP model of movement is developed based on Tsallis divergence regularization. Section 4 introduces the two new dissimilarity measures between nodes. Illustrative examples and experiments on node clustering and semi-supervised classification tasks are presented in Section 5. Finally, Section 6 concludes the work.

---

<sup>4</sup>For an undirected graph and up to a scaling factor.

## 2 The randomized shortest paths framework

This section provides a short summary of the standard randomized shortest paths framework, based on the KL divergence, before introducing the alternative formulation (Subsection 2.3) and then developing its sparse version (Section 3).

### 2.1 Background and notation

Let us first introduce some background and notation [34, 36]. Consider a weighted directed<sup>5</sup>, strongly connected, graph,  $G = (\mathcal{V}, \mathcal{E})$ , with a set  $\mathcal{V}$  of  $n$  nodes and a set  $\mathcal{E}$  of directed edges. The directed edge connecting node  $i$  to node  $j$  is denoted by  $(i, j)$ . Moreover, the *adjacency matrix*  $\mathbf{A} = (a_{ij}) \geq 0$  of  $G$  contains directed affinities between nodes. We further assume that there are no self-loops in the network, that is,  $a_{ii} = 0$  for all  $i$  (a simple graph).

A natural random walk on the graph is defined from this adjacency matrix in the usual way. The *reference transition probabilities* associated to each node are set proportionally to the affinities of the incident edges and then normalized in order to sum to one. Alternatively, they can also be chosen as uniform, depending on the problem,

$$p_{ij}^{\text{ref}} = \frac{a_{ij}}{\sum_{j'=1}^n a_{ij'}} \quad (\text{natural random walk}) \quad \text{or} \quad p_{ij}^{\text{ref}} = \frac{1}{|\text{Succ}(i)|} \quad (\text{uniform distribution}) \quad (1)$$

where  $\text{Succ}(i)$  is the set of successor nodes of node  $i$  and  $|\text{Succ}(i)|$  its cardinality. The matrix  $\mathbf{P}_{\text{ref}} = (p_{ij}^{\text{ref}})$  is stochastic and is called the transition matrix of the reference random walk on the graph.

Furthermore, a transition *cost*,  $c_{ij} \geq 0$ , is associated to each edge  $(i, j)$  of the network  $G$  and the resulting cost matrix is defined as  $\mathbf{C} = (c_{ij})$ . If there is no edge linking  $i$  to  $j$  ( $a_{ij} = 0$ ), the cost is assumed to take a very large value and the product  $a_{ij}c_{ij} = 0$  by convention. Usually, costs are set independently of the affinities, depending on the application, but, if there are no obvious costs associated to the problem, we can, e.g., set  $c_{ij} = 1/a_{ij}$  as in electric networks (see [36] for a discussion).

A path  $\varphi$  is a finite sequence of transitions to adjacent nodes on  $G$  (including cycles), initiated from a source node  $s$  and stopping in some target node  $t$ . This target node is transformed into an *absorbing* and *killing* node. That is, when this node is reached, the random walkers stops his walk and disappears. This means that the corresponding row  $t$  of the transition probabilities matrix  $\mathbf{P}_{\text{ref}}$  is set to zero – the matrix therefore becomes sub-stochastic and represents a killed random walk on  $G$ . The *total cost* of a path,  $\tilde{c}(\varphi)$ , is simply the sum of the edge costs  $c_{ij}$  along  $\varphi$  while the *length* of a path  $\ell(\varphi)$  (or simply  $\ell$ ) is the number of steps needed for following that path from  $s$  to  $t$ . Finally, the set of all paths connecting  $s$  to  $t$  is denoted by  $\mathcal{P}_{st}$ .

### 2.2 The standard randomized shortest paths framework

The *randomized shortest paths* (RSP) framework [99], coming from transportation sciences [2], was further developed by exploiting basic concepts from statistical physics [36, 63, 99, 124]. A similar model was proposed in reinforcement learning and process control [113, 114], from a different

---

<sup>5</sup>If the graph is undirected, it is assumed that each undirected edge is composed of two directed edges with the same weight in the two opposite directions (reciprocal edges).

perspective. The RSP is based on a system of full paths (the “macro” level – a bag of paths connecting the source node to the target node<sup>6</sup>) instead of standard “local” flows defined on nodes and edges (the “micro” level) [1, 28]. Among others, it provides an optimal, randomized, routing policy from source node  $s$  to target node  $t$  based on the minimization over the set of paths  $\varphi \in \mathcal{P}_{st}$  of the (relative) KL-based free energy of statistical physics [56, 92, 97],

$$\left\{ \begin{array}{l} \text{minimize}_{\{P(\varphi)\}_{\varphi \in \mathcal{P}_{st}}} \quad \phi_{st}^{\text{KL}}(P) = \underbrace{\sum_{\varphi \in \mathcal{P}_{st}} P(\varphi) \tilde{c}(\varphi)}_{\text{expected cost}} + T \underbrace{\sum_{\varphi \in \mathcal{P}_{st}} P(\varphi) \log\left(\frac{P(\varphi)}{\tilde{\pi}(\varphi)}\right)}_{\text{KL regularization}} \\ \text{subject to} \quad \sum_{\varphi \in \mathcal{P}_{st}} P(\varphi) = 1 \end{array} \right. \quad (2)$$

where  $\tilde{c}(\varphi) = \sum_{\tau=1}^{\ell} c_{s(\tau-1)s(\tau)}$  is the total cumulated cost along path  $\varphi$  when visiting the sequence of nodes, or states,  $(s(\tau))_{\tau=0}^{\ell}$  in the sequential order and  $\ell$  is the length of path  $\varphi$ . Furthermore,  $\tilde{\pi}(\varphi) = \prod_{\tau=1}^{\ell} p_{s(\tau-1)s(\tau)}^{\text{ref}}$  is the random walk probability of the path, that is, the product of the reference transition probabilities (1) along hitting path  $\varphi$  ending in (killing and absorbing) target node  $t$ . The objective function is a mixture of two dissimilarity terms with the temperature  $T > 0$  balancing the trade-off between the two quantities. The first term is the expected cost for reaching target node from source node (favoring shorter paths – *exploitation*). The second term corresponds to the Shannon relative entropy, or Kullback-Leibler divergence, between the path probability distribution and the reference path probability distribution (introducing randomness – *exploration*). For a low temperature  $T$ , shorter paths are favored whereas when  $T$  is large, paths are chosen according to their probability in the reference random walk on  $G$  (see Equation (1)).

As well-known, this free energy minimization problem, akin to maximum entropy models [24, 56, 60], leads to an *Gibbs-Boltzmann* probability distribution on the set of paths (see, e.g., [36]),

$$P^*(\varphi) = \frac{\tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)]}{\sum_{\varphi' \in \mathcal{P}_{st}} \tilde{\pi}(\varphi') \exp[-\theta \tilde{c}(\varphi')]} = \frac{\tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)]}{\mathcal{Z}} \quad (3)$$

where  $\theta = 1/T$  is the inverse temperature and the denominator  $\mathcal{Z} = \sum_{\varphi \in \mathcal{P}_{st}} \tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)]$  is the *partition function* of the system. This provides the (randomized) *optimal policy* in terms of paths to follow at the “macro” (paths) level, that is, the *optimal* probability distribution over all possible paths from  $s$  to  $t$ .

It can further be shown that this probability distribution defines a local, optimal, policy which turns out to be a Markov chain at the “micro” (node) level. In this context, the optimal transition probabilities of following any edge  $(i, j)$  (the “local policy”) induced by the set of paths  $\mathcal{P}_{st}$  and their probability mass (3) are<sup>7</sup>

$$p_{ij}^* = p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \phi_{jt}^{\text{KL}}(P^*) - \phi_{it}^{\text{KL}}(P^*))] \text{ for all } i \neq t \quad (4)$$

and  $p_{tj}^* = 0$  for target node  $t$  and all  $j$ . The free energy between  $i$  and  $t$  at the optimal probability distribution<sup>8</sup>,  $\phi_{it}^{\text{KL}}(P^*)$ , can be computed by solving a system of linear equations of size  $n$  [36, 63, 99]. As already mentioned, it defines a *biased* random walk on  $G$  where the random walker is more and

<sup>6</sup>For a generalization to several input-outputs, see [46].

<sup>7</sup>Here, we used  $\phi_{it}^{\text{KL}}(P^*) = -\frac{1}{\theta} \log[z_{it}]$  and  $p_{ij}^* = p_{ij}^{\text{ref}} \exp[-\theta c_{ij}] z_{jt} / z_{it}$  (see [36, 63, 99] for details).

<sup>8</sup>Providing the directed free energy distance, see [63].

more “attracted” by the target node  $t$  when  $T$  decreases. Interestingly, the policy is independent of the source node  $s$ . It also implies that for any path  $\varphi$  visiting nodes  $s(\tau)$ ,  $\tau = 0, \dots, \ell(\varphi)$ , we have  $P^*(\varphi) = \prod_{\tau=1}^{\ell} p_{s(\tau-1)s(\tau)}^*$ .

### 2.3 An alternative form for the randomized shortest paths

The previous path-based objective function (2) can be transformed into a “micro” form based on local flows (see [3, 7, 99]) which will be used later for deriving the sparse RSP. In this new form, one possible way to compute the policy is by sequentially solving the primal and the dual problems. In the case of KL divergence regularization, the algorithm cannot compete against standard procedures developed in [36, 63, 99], which are quite efficient. However, the advantage is that this algorithm can be easily adapted to other divergence regularizations.

As shown in A.1, the equivalent problem at the local level aims at minimizing the free energy with respect to the transition probabilities instead of paths probabilities,

$$\begin{array}{l}
 \left. \begin{array}{l}
 \text{minimize } \phi_{st}^{\text{KL}}(\mathbf{P}) = \sum_{(i,j) \in \mathcal{E}} \bar{n}_i p_{ij} \left( c_{ij} + T \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right) \\
 \text{subject to } \bar{n}_j = \sum_{i \in \mathcal{P}red(j)} \bar{n}_i p_{ij} + \delta_{sj}, \text{ for all nodes } j \in \mathcal{V} \\
 \sum_{j \in \mathcal{S}ucc(i)} p_{ij} = 1 \text{ for all } i \neq t \\
 p_{tj} = 0 \text{ for absorbing and killing node } t \text{ and all } j \in \mathcal{V} \\
 \mathbf{P} \geq \mathbf{0} \\
 \bar{\mathbf{n}} \geq \mathbf{0}
 \end{array} \right\} \quad (5)
 \end{array}$$

where  $\mathcal{P}red(j)$  is the set of predecessor nodes of node  $j$ . This formulation of the optimization problem is derived from Equation (22), A.1, and then solved in A.2. In this Equation (5), the  $p_{ij}$  are the elements of the transition matrix  $\mathbf{P}$  to be found (the *routing policy*), the  $p_{ij}^{\text{ref}}$  are the elements of the transition matrix of the reference random walk on the graph (see Equation (1)), and the  $\bar{n}_j$  are the expected numbers of visits to the nodes when walking on  $G$  according to the transition matrix  $\mathbf{P}$ , provided by  $\bar{n}_j = \sum_{i \in \mathcal{P}red(j)} \bar{n}_i p_{ij} + \delta_{sj}$  for nodes in  $\mathcal{V}$  (see, e.g., [90, 109]). Note that because the input flow in source node  $s$  is 1 and target node  $t$  is the only absorbing and killing node in the network (the sink),  $\bar{n}_t = 1$ .

Following the Lagrange formulation of the problem appearing in 23 and inspired by discrete space-time optimal control (see, e.g., [76]), the  $p_{ij}$  and the  $\bar{n}_j$  can be considered as independent (the relations between these variables are encoded in the (linear) constraints). Notice that it is not necessary to impose  $p_{ij} = 0$  for missing edges as this will be enforced automatically by the algorithm thanks to the use of the KL divergence.

The intuition behind the Equation (5) is as follows. The first line (objective function) computes the total expected cost plus KL divergence when following edge  $(i, j)$  because  $\bar{n}_i p_{ij} = \bar{n}_{ij}$  represents the flow in edge  $(i, j)$  (expected number of visits to  $i$  times the probability of following  $(i, j)$ ). The second line (first constraint) provides the expression for computing the expected number of visits to each node. The last lines state the sum-to-one as well as the non-negativity constraints, which are in fact not necessary in the case of a KL regularization but which will be important in the next section when dealing with Tsallis divergence.

In A (see Equations (24) and (26)), it is shown that the transition probabilities, providing the routing policy, can be computed by sequentially updating the Lagrange parameters  $\lambda_i^{\text{KL}}$  (associated to the computation of the expected number of visits to nodes) from the transition probabilities and vice-versa thanks to

- Compute the Lagrange parameters by solving the system of linear equations:

$$\lambda_i^{\text{KL}} - \sum_{j \in \text{Succ}(i)} p_{ij} \lambda_j^{\text{KL}} = \sum_{j \in \text{Succ}(i)} p_{ij} \left( c_{ij} + T \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right) \quad \text{for all } i \in \mathcal{V} \quad (6)$$

- Compute the transition probabilities:

$$p_{ij} = \begin{cases} \frac{p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \lambda_j^{\text{KL}})]}{\sum_{k \in \text{Succ}(i)} p_{ik}^{\text{ref}} \exp[-\theta(c_{ik} + \lambda_k^{\text{KL}})]} & \text{for all } (i, j) \in \mathcal{E} \\ 0 & \text{for the absorbing and killing target node } i = t \end{cases} \quad (7)$$

Initially (at iteration 0), the transition probabilities are set to the reference transition probabilities,  $p_{ij} = p_{ij}^{\text{ref}}$ .

Note that because the target node  $t$  is absorbing and killing, Equation (6) provides  $\lambda_t^{\text{KL}} = 0$ . We also observe that, for missing edges, the transition probabilities are equal to zero, as it should be. Interestingly, after convergence, the Lagrange parameters are nothing else than the *optimal directed free energy distance*,  $\lambda_i^{\text{KL}} = \phi_{it}^{\text{KL}}(\mathbf{P}^*)$ , appearing in Equation (2) and evaluated at the Gibbs-Boltzmann distribution  $\mathbf{P}^*$  (Equation (3)), as discussed in A.2. Moreover, notice that, for an undirected graph and edge costs defined as  $c_{ij} = 1/a_{ij}$ , the Equation (6) is nothing more than the harmonic function (induced by Kirchhoff's current law and Ohm's law) computing the voltage associated to the nodes, when considering sources of voltage on nodes [29].

We now follow the same derivation with the difference that we will be using Tsallis instead of KL divergence for defining a sparse policy and the corresponding (directed) FE distance.

### 3 Sparse randomized shortest paths

We now turn to the development of the main contribution of the paper, the introduction of the Tsallis RSP routing framework. Depending on  $T$ , it provides a kind of “compressed graph structure” keeping only the most relevant edges for communicating efficiently from source node  $s$  to target node  $t$  (see Figure 2 for an illustration).

#### 3.1 Statement of the problem

If we have a set of  $m$  mutually exclusive random outcomes of a chance experiment whose probabilities  $p_j \geq 0$  sum to one, as well as known reference probabilities  $p_j^{\text{ref}} \geq 0$ , the Tsallis ([115, 116] also called Havrda-Charvat [50, 60]) directed  $r$ -divergence between the two probability distributions is given by

$$H_r(\mathbf{P}|\mathbf{P}^{\text{ref}}) = \frac{1}{r-1} \left( \sum_{j=1}^m \left( \frac{p_j^r}{(p_j^{\text{ref}})^{r-1}} \right) - 1 \right) = \frac{1}{r-1} \sum_{j=1}^m p_j \left( \left( \frac{p_j}{p_j^{\text{ref}}} \right)^{r-1} - 1 \right) \quad (8)$$



where, in this work, the parameter  $r$  will be assumed to be larger than one,  $r > 1$ , but the most common value is  $r = 2$ . The role of  $r$  will be discussed after Equation (12). This measure (8) generalizes the KL divergence [60] in the sense that it converges to this quantity when  $r \rightarrow 1^+$ . It is also closely related to the Simpson-Gini index widely used in decision trees, as well as other measures of uncertainty and diversity (see, e.g., [60, 62]). As for the more standard KL divergence, this measure is non-negative and quantifies the divergence between the two probability distributions.

The Tsallis divergence will now be used in order to define an optimal, sparse, policy. From Equation (5), the problem aims to minimize the Tsallis-based free energy function instead of the KL-based free energy of Equation (2),

$$\begin{array}{l}
 \left. \begin{array}{l}
 \text{minimize}_{\mathbf{P}} \quad \phi_{st}^{\text{TS}}(\mathbf{P}) = \sum_{(i,j) \in \mathcal{E}} \bar{n}_i p_{ij} \left( c_{ij} + \frac{T}{r-1} \left( \left( \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right)^{r-1} - 1 \right) \right) \\
 \text{subject to} \quad \bar{n}_j = \sum_{i \in \mathcal{P}red(j)} \bar{n}_i p_{ij} + \delta_{sj}, \text{ for all nodes } j \in \mathcal{V} \\
 \sum_{j \in \mathcal{S}ucc(i)} p_{ij} = 1 \text{ for all } i \neq t \\
 p_{tj} = 0 \text{ for absorbing and killing node } t \text{ and all } j \in \mathcal{V} \\
 \mathbf{P} \geq \mathbf{0} \\
 \bar{\mathbf{n}} \geq \mathbf{0}
 \end{array} \right\} \quad (9)
 \end{array}$$

For convenience, let us renumber the nodes in such a way that node 1 is the source node and node  $n$  (last node) is the target node<sup>9</sup>. The Lagrange function (see A.2, especially Equation (23), for details) integrating the equality constraints (but not the non-negativity constraints restricting the domain of  $\mathbf{P}$  and  $\bar{\mathbf{n}}$ ) becomes

$$\begin{aligned}
 \mathcal{L}(\mathbf{P}, \bar{\mathbf{n}}; \boldsymbol{\mu}, \boldsymbol{\lambda}_{\text{TS}}) &= \sum_{i \in \mathcal{V} \setminus n} \bar{n}_i \sum_{j \in \mathcal{S}ucc(i)} p_{ij} \left( c_{ij} + \frac{T}{r-1} \left( \left( \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right)^{r-1} - 1 \right) \right) \\
 &+ \sum_{j \in \mathcal{V}} \lambda_j^{\text{TS}} \left( \sum_{i \in \mathcal{P}red(j)} \bar{n}_i p_{ij} + \delta_{1j} - \bar{n}_j \right) + \sum_{i \in \mathcal{V} \setminus n} \mu_i \left( 1 - \sum_{j \in \mathcal{S}ucc(i)} p_{ij} \right) \quad (10)
 \end{aligned}$$

where we also have to deal with non-negativity constraints on the transition probabilities,  $p_{ij} \in \mathbb{R}_+$ , and the expected number of visits,  $\bar{n}_i \in \mathbb{R}_+$ . As for the standard RSP, for  $\theta = 1/T \rightarrow 0^+$ , this model becomes a simple absorbing random walk on  $G$  with transition probabilities provided by Equation (1) (reference probabilities). For  $\theta \rightarrow \infty$ , it provides least-cost routing concentrated on shortest paths.

Let us proceed like in previous section for optimizing the objective function with respect to the transition probabilities. As before, it is not necessary to impose  $p_{ij} = 0$  for missing edges because these constraints will be satisfied automatically when recomputing the transition probabilities.

But first, a remark concerning related work before proceeding. If we compare the objective function used in (10) with the one studied in [73, 74] (see Equation (18) of [74] – also leading to a sparse policy), because the flow in edge  $(i, j)$  is provided by  $\bar{n}_{ij} = \bar{n}_i p_{ij}$ , we observe that they are quite different. Indeed, the objective function of [73, 74] can be rewritten in our notation as  $c'_{ij} = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{S}ucc(i)} \bar{n}_i p_{ij} (c_{ij} + \gamma \bar{n}_i p_{ij})$  where  $\gamma \geq 0$  is a parameter controlling sparseness. Note

---

<sup>9</sup>It is assumed that  $s \neq t$ .

also that [73, 74] consider a symmetric cost matrix and net flows, instead of raw flows in the present paper.

As for previous section, and as detailed in A, our optimization procedure optimizes sequentially the objective function by Lagrange duality [25, 43, 83] and a variant of the Arrow-Hurwicz-Uzawa algorithm [5]. More precisely, the Lagrange function is first minimized with respect to the transition probabilities  $p_{ij}$  subject to their constraints while considering the Lagrange parameters  $\lambda_i^{\text{TS}}$  as fixed. Then, Lagrange parameters are computed by maximizing the dual problem. The two steps are iterated until convergence to a stationary point of the objective function (9), which is guaranteed because each sub-problem reaches its optimum uniquely [9]. A discussion of the convexity of the objective function with respect of the edge flows appears in B. In short, it is shown by heuristic arguments that the objective function appearing in Equation (9) is convex with respect to the edge flows  $\bar{n}_{ij}$ , although a formal proof is left for further work. Then, by using the same reasoning as for the KL divergence case of Subsection A.1, convergence to a global minimum should be guaranteed. In practice, for all our experimental runs, we observed that the duality gap is equal to zero, showing that a global minimum is reached.

Notice that there is an important difference with the previous section (KL regularization), namely that the non-negativity of the transition probabilities is no more guaranteed. We therefore have to deal with these inequality constraints and rely on the Karush-Kuhn-Tucker conditions.

### 3.2 Computation of the transition probabilities

The Lagrange function will first be minimized with respect to the transition probabilities (subject to their constraints), with the Lagrange parameters fixed. To this end, let us re-arrange the Lagrange function (10) by gathering all the terms depending on the transition probabilities,

$$\begin{aligned}
& \mathcal{L}(\mathbf{P}, \bar{\mathbf{n}}; \boldsymbol{\mu}, \boldsymbol{\lambda}_{\text{TS}}) \\
&= \sum_{i=1}^{n-1} \bar{n}_i \left( \sum_{j=1}^n p_{ij} \underbrace{(c_{ij} + \lambda_j^{\text{TS}})}_{\text{augmented costs } \mathbf{c}'_{ij}} + \underbrace{\frac{T}{r-1} \sum_{j=1}^n p_{ij} \left( \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right)^{r-1}}_{\text{regularization term}} \right) + \sum_{i=1}^{n-1} \mu_i \left( 1 - \sum_{j=1}^n p_{ij} \right) \\
&\quad - (\bar{\mathbf{n}}_{\bullet} - 1) + \sum_{i=1}^n \sum_{j=1}^n \lambda_j^{\text{TS}} (\delta_{1j} - \bar{n}_j)
\end{aligned} \tag{11}$$

because  $\bar{n}_n = 1$  and  $\bar{\mathbf{n}}_{\bullet} = \sum_{i=1}^n \bar{n}_i$ . In this last expression, we defined the *augmented costs* for each row  $i$  (and thus each transient node  $i$ ), considered as a column vector,  $\mathbf{c}'_i = \mathbf{row}_i(\mathbf{C}) + \boldsymbol{\lambda}_{\text{TS}}$ .

From (11), we observe that the part of the objective function depending on the transition probabilities (first line) is a sum of  $(n-1)$  terms that can be optimized independently at the level of each node and its incident links. The same is true for the constraints (sum-to-one and non-negativity) which are also operating at the level of the nodes. Moreover, the expected numbers of visits  $\bar{n}_i$  are considered as independent from the transition probabilities because of the Lagrangian formulation of the problem.

Therefore, the problem that needs to be solved for each transient node  $i = 1, \dots, (n-1)$  in turn

is

$$\left\{ \begin{array}{l} \underset{\mathbf{p}_i}{\text{minimize}} \quad \underbrace{(\mathbf{c}'_i)^T \mathbf{p}_i}_{\text{expected cost}} + \underbrace{\frac{T}{r-1} \mathbf{p}_i^T (\mathbf{p}_i \div \mathbf{p}_i^{\text{ref}})^{(r-1)}}_{r\text{-power regularization term}} \\ \text{subject to} \quad \mathbf{e}^T \mathbf{p}_i = 1 \\ \mathbf{p}_i \geq \mathbf{0} \end{array} \right. \quad (12)$$

together with

$$\left\{ \begin{array}{l} \mathbf{c}'_i = \mathbf{row}_i(\mathbf{C}) + \lambda_{\text{TS}} \\ \mathbf{p}_i^{\text{ref}} = \mathbf{row}_i(\mathbf{P}_{\text{ref}}) \\ r > 1 \end{array} \right.$$

where  $\div$  is the elementwise division,  $(r-1)$  is the elementwise power and  $T = 1/\theta > 0$  monitors the balance between expected augmented cost and Tsallis divergence minimization. This shows that the Lagrange parameters  $\lambda_j^{\text{TS}}$  can be interpreted as the virtual costs that have to be added to the real costs  $c_{ij}$  in order to be able to compute the local transition probabilities at the level of each node thanks to Equation (12).

Note that the  $r$  parameter controls the impact of the costs on the probabilities. For example, for uniform reference probabilities and linear costs,  $\mathbf{c} = [1, 2, 3, 4, 5]^T$ , the form of the resulting probability distribution will be decreasing and convex when  $r < 2$  ( $\mathbf{p} = [0.480, 0.295, 0.156, 0.0602, 0.00927]^T$  when  $r = 1.5$  and  $T = 1$ ), linear when  $r = 2$  ( $\mathbf{p} = [0.4, 0.3, 0.2, 0.1, 0.0]^T$  with  $T = 1$ ), and concave when  $r > 2$  ( $\mathbf{p} = [0.289, 0.262, 0.229, 0.182, 0.0375]^T$  when  $r = 4$  and  $T = 1$ ). On the other hand, the temperature  $T$  mainly controls the sparseness of the distribution ( $\mathbf{p} = [0.533, 0.333, 0.133, 0.0, 0.0]^T$  when  $r = 2$  with  $T = 0.5$  and  $\mathbf{p} = [0.240, 0.220, 0.200, 0.180, 0.160]^T$  when  $r = 2$  with  $T = 5$ ).

Thus, the problem of computing the transition probabilities reduces to the minimization of a linear function with a  $r$ -power regularization term on the probability simplex. Interestingly, for a uniform reference distribution, this problem of minimizing a linear objective function with quadratic as well as  $r$ -power regularization has been studied in the fuzzy clustering literature [57, 85]. In this context, the objective was to fuzzify the membership functions thanks to a quadratic regularization, similar to the Tsallis entropy, in a fuzzy  $k$ -means. Note that in their original work, the constant term  $-1$  (minus one) present in the Tsallis entropy is not taken into account – the authors simply consider a quadratic regularization, which has no effect on the solution. Note that [85] solves the quadratic regularization problem with  $r = 2$  while [57] provides an algorithm for the more general case  $r > 1$ .

Inspired by this previous work, the problem (12) is solved by a procedure (spmin) computing a sparse minimum derived in C for each node  $i \neq t$ ,

$$\mathbf{p}_i^* = \text{spmin}(\mathbf{c}'_i, \mathbf{p}_i^{\text{ref}}, r, T) \quad \text{for } i = 1, \dots, (n-1) \quad (13)$$

which returns a possibly sparse discrete probability distribution  $\mathbf{p}_i^*$ . In short, the solution for the general case is of the form (Equation (49))

$$\mathbf{p}_i^* = \mathbf{p}_i^{\text{ref}} \circ \sqrt[r-1]{\frac{r-1}{rT} [\mu_i \mathbf{e} - \mathbf{c}'_i]_+} \quad (14)$$

where  $[x]_+ = \max(x, 0)$ : if  $x$  is negative, it is put to 0, and  $\circ$  is the elementwise Hadamard product. The vector  $\mathbf{e}$  is a column vector full of 1's and the threshold  $\mu_i$  must be chosen in order to satisfy the sum-to-one constraint,  $(\mathbf{p}_i^*)^T \mathbf{e} = 1$ . The expression (14) is the counterpart of Equation (4) based on the KL divergence.

The procedure (13) is run on every node  $i \neq n$  in order to obtain an updated transition matrix  $\mathbf{P}$ . Again, the transition probabilities only depend on the Lagrange parameters  $\boldsymbol{\lambda}_{\text{Ts}}$  through the augmented costs.

Based on previous work [57, 85], three algorithms are developed in the C. More precisely, in C.1, we consider the special case of a quadratic regularization term,  $r = 2$ , whereas the general case  $r > 1$  is developed in C.2. This is because the quadratic case is simpler and leads to an efficient algorithm based on a linear search once the nodes have been sorted by increasing cost, while the general case is handled by using a bisection search which turns out to be slower in practice. Our contribution with respect to [57, 85] is the introduction of non-uniform reference probabilities in C.3 which allow to deal with the Tsallis divergence instead of the Tsallis entropy. Let us now turn to the computation of the Lagrange parameters.

### 3.3 Computation of the Lagrange parameters

The second step, i.e. the computation of the Lagrange parameter vector  $\boldsymbol{\lambda}_{\text{Ts}}$ , follows the same principle as for the KL divergence (see Equation (6) and Subsection A.2 for details), the only difference being the definition of the divergence. Indeed, the elements of  $\tilde{\mathbf{h}}_{\text{KL}}$ , based on the KL divergence, in Equation (25), now become  $\tilde{h}_i^{\text{Ts}} = \frac{1}{r-1} \sum_{j \in \text{Succ}(i)} p_{ij} \left( (p_{ij}/p_{ij}^{\text{ref}})^{r-1} - 1 \right)$  for the Tsallis divergence.

By similitude with the KL regularization (see Equation (6)), we define the *Tsallis directed free energy* potential as the values of the Lagrange parameters obtained after convergence of the optimization problem,

$$\boldsymbol{\phi}_{\text{Ts}} \triangleq \boldsymbol{\lambda}_{\text{Ts}} \quad (15)$$

and we verified that they correspond to the minimized free energy objective function (9). These quantities are obtained by solving the following system of linear equations

$$\lambda_i^{\text{Ts}} - \sum_{j \in \text{Succ}(i)} p_{ij} \lambda_j^{\text{Ts}} = \sum_{j \in \text{Succ}(i)} p_{ij} \left[ c_{ij} + \frac{T}{r-1} \left( \left( \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right)^{r-1} - 1 \right) \right] \quad \text{for all } i \in \mathcal{V} \quad (16)$$

with respect to  $\boldsymbol{\lambda}_{\text{Ts}}$ . Note that these equations imply  $\lambda_t^{\text{Ts}} = 0$  for target node  $t$ . In matrix form, we have

$$(\mathbf{I} - \mathbf{P})\boldsymbol{\lambda}_{\text{Ts}} = \tilde{\mathbf{c}} + T\tilde{\mathbf{h}}_{\text{Ts}} \quad (17)$$

where  $\tilde{\mathbf{c}} = (\mathbf{P} \circ \mathbf{C})\mathbf{e}$  and  $\tilde{\mathbf{h}}_{\text{Ts}} = \frac{1}{r-1} [\mathbf{P} \circ ((\mathbf{P} \div \mathbf{P}_{\text{ref}})^{(r-1)} - \mathbf{E})]\mathbf{e}$ ,  $\mathbf{e}$  is a column vector containing 1 on each row,  $\mathbf{E}$  is a square matrix full of 1's,  $\circ$  is the elementwise matrix product, and  $(r-1)$  is the elementwise power.

Finally, after initializing the transition probabilities to the reference probabilities, the overall procedure for computing the optimal randomized policy based on Tsallis divergence aims at iterating Equations (17) and (13) until convergence. Each of the two steps has a unique optimal solution so that the iterative procedure converges [9]. As before, we observed that, in all our runs, the duality gap is always zero, showing that a global minimum is reached. After convergence, the algorithm provides an optimal, possibly sparse, routing policy taking the form of Markov chain with absorbing, killing, node  $n$  (or  $t$  before renumbering the nodes).

## 4 Derived dissimilarities

We now define two new dissimilarity measures between nodes interpolating between the least cost and the commute-cost<sup>10</sup>, up to a constant scaling factor. These quantities are the counterparts of the RSP dissimilarity and the FE distance based on KL regularization. As for the KL divergence [36, 63, 124], the quantity  $\phi_{it}^{\text{TS}}$ , provided by the Lagrange parameter  $\lambda_i^{\text{TS}}$  after convergence (see Equation (15)), is called the Tsallis directed FE from node  $i$  to absorbing, killing, node<sup>11</sup>  $t$ . Accordingly, the *Tsallis FE distance* is the symmetrized quantity

$$\Delta_{st}^{\text{TSFE}} = \frac{\phi_{st}^{\text{TS}} + \phi_{ts}^{\text{TS}}}{2} \quad (18)$$

Interestingly, we found that this quantity satisfies the triangle inequality on all the investigated datasets and values of the  $\theta$  parameter (see the next, experimental, section). We therefore conjecture that it defines a distance measure between nodes (as in the case for the KL divergence [36]) and hope to prove it in future work.

Moreover, the directed Tsallis RSP dissimilarity is based on the total expected cost for reaching target node  $t$  from node  $s$  when following the optimal policy, which is given by

$$\langle c \rangle_{st} = \sum_{(i,j) \in \mathcal{E}} \bar{n}_i p_{ij} c_{ij} = \sum_{(i,j) \in \mathcal{E}} \bar{n}_{ij} c_{ij} \quad (19)$$

where  $p_{ij}$  is computed by Equation (13). The *Tsallis randomized shortest paths dissimilarity* is directly deduced from the previous expression,

$$\Delta_{st}^{\text{TSRSP}} = \langle c \rangle_{st} + \langle c \rangle_{ts} \quad (20)$$

This dissimilarity, however, does not satisfy the triangle inequality for all values of the parameter  $\theta$ ; we are thus in the same situation as for the KL regularization: the Tsallis randomized shortest paths dissimilarity is not a distance measure. The FE distance and the RSP dissimilarity will be compared on pattern recognition tasks in the next section.

## 5 Experiments

In this section, we first present an illustrative example of the sparsity of the routing policy induced by the Tsallis-based RSP, by visualizing the net flows on a weighted undirected graph. For each edge  $(i, j)$ , the *net flow* is defined as  $net_{ij} = \max(\bar{n}_{ij} - \bar{n}_{ji}, 0)$  and is oriented in the direction of the positive flow (at most one per edge). Then, we evaluate our methods by comparing them with other state-of-the-art dissimilarity measures on node clustering and semi-supervised classification tasks<sup>12</sup>. We have to stress that our goal here is not to propose new node clustering or semi-supervised classification algorithms outperforming state-of-the-art techniques. Rather, the aim is to investigate if the Tsallis-based RSP model is able to capture the community structure of networks in an accurate way, compared to other, existing, state-of-the-art dissimilarity measures between nodes.

<sup>10</sup>And thus also the resistance distance based on the effective resistance when  $c_{ij}$  is defined as  $1/a_{ij}$ , like in electrical networks [33]. This property only holds in the case of an undirected graph

<sup>11</sup>Recall that the target node is always transformed into an absorbing and killing node.

<sup>12</sup>Note that all results were obtained with Matlab (version R2019) running on an Intel Xeon with  $2 \times 8$  core processors of 3.6 GHz and 128 GB of RAM.

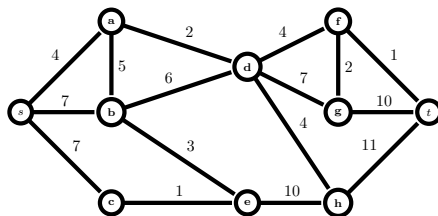


Figure 1: A small undirected graph composed of 10 nodes with costs on edges; adapted from [94].

## 5.1 Illustrative example

Consider the graph represented in Figure 1, containing a source node  $s$  and a target node  $t$ . The weight on the edges represents the cost associated to the transitions.

We illustrate how the parameter  $\theta$  influences the sparsity of the transition probabilities (the randomized routing policy) by representing the net flow from the source to the target. These flows are depicted for the value  $r = 2$  in Figure 2, where only positive net flows are drawn.

As the value of  $\theta$  increases, some edges gradually become unused (the flow in the edge is equal to zero). Eventually, the flow is entirely concentrated on the shortest path (already when  $\theta = 2$  in our example). This clearly shows that the RSP routing policy becomes gradually sparser when the parameter  $\theta$  increases. This property is also observed for other values of the parameter  $r$ .

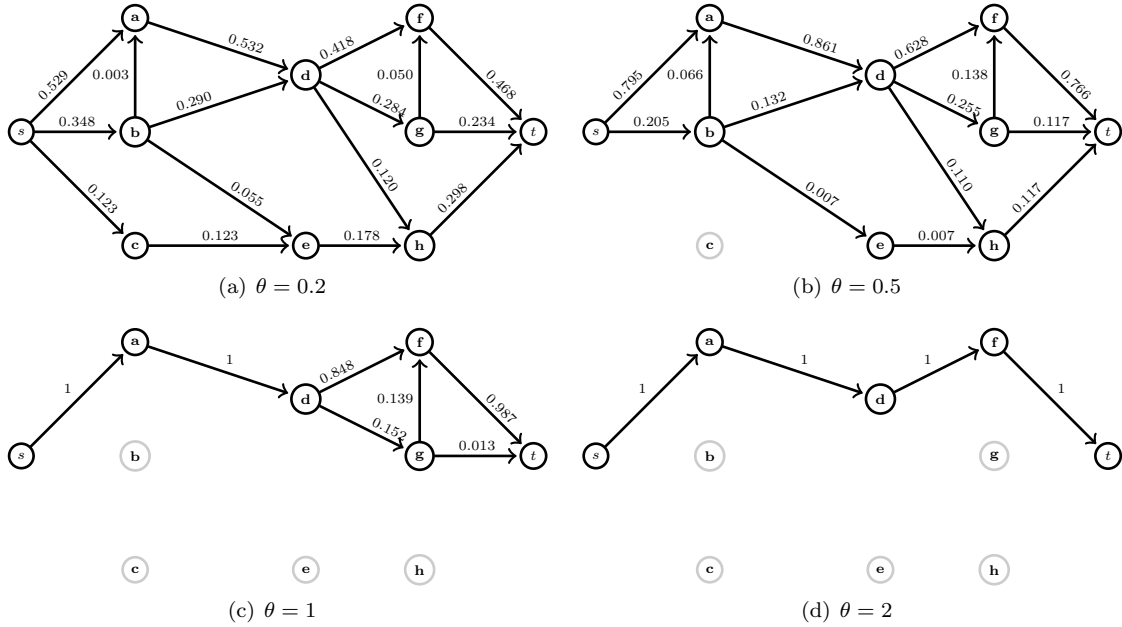
## 5.2 Global experimental setup

The Tsallis FE (FETsallis) and the Tsallis RSP (RSPTsallis) dissimilarities will be assessed in two different contexts: first, a node clustering task and, second, a graph-based semi-supervised classification task aiming to categorize unlabeled nodes. For all RSP-based methods, the reference transition probabilities are set to  $p_{ij}^{\text{ref}} = a_{ij} / \sum_{j'=1}^n a_{ij'}$ , corresponding to a natural random walk on the graph (see Equation (1)). In addition, the costs on the edges are defined as  $c_{ij} = 1/a_{ij}$ , as for electrical networks. As part of the experiments, four dissimilarity matrices between nodes as well as four kernels on a graph will be used as baseline methods to assess our methods.

### Baseline dissimilarities between nodes

- ▶ The **Free Energy** distance (FE, based on KL divergence) and the **Randomized Shortest Paths Dissimilarity** (RSP, also based on KL divergence) depending on an inverse temperature parameter  $\theta = 1/T$ . Already presented earlier in Section 2, these methods have been shown to perform well in node clustering [102] as well as in semi-supervised classification tasks [36].
- ▶ The **Logarithmic Forest** distance (LF). Introduced in [16], it relies on the matrix-forest theorem [19] and defines a family of distances interpolating (up to a scaling factor) between the shortest-path distance and the resistance distance, depending on a parameter  $\alpha$ .
- ▶ The **Shortest Path** distance (SP). The distance corresponds to the cost along the shortest path between two nodes  $i$  and  $j$ , derived from the cost matrix  $\mathbf{C}$ .

These dissimilarity matrices are transformed into inner products (kernel matrices) by classical multidimensional scaling (see later).



**Figure 2:** Representation of the net flow from  $s$  to  $t$  in function of  $\theta$  for the Tsallis RSP and using  $r = 2$ .

### Baseline kernels on a graph

- The **Neumann** kernel [100] (Katz), initially proposed in [61] as a method of computing similarities, and defined as  $\mathbf{K} = (\mathbf{I} - \alpha\mathbf{A})^{-1} - \mathbf{I}$ . The  $\alpha$  parameter has to be chosen positive and smaller than the inverse of the spectral radius of  $\mathbf{A}$ ,  $\rho(\mathbf{A}) = \max_i(|\lambda_i|)$ .
- The **Logarithmic Communicability** kernel (lCom) proposed in [55] as the logarithmic version of the exponential diffusion kernel [66], also known as the communicability measure [32],  $\mathbf{K} = \ln(\text{expm}(t\mathbf{A}))$ ,  $t > 0$ , where  $\text{expm}$  is the matrix exponential.
- The **Sigmoid Commute Time** kernel. Proposed in [122], it is obtained by applying a sigmoid transform [100] on the commute time kernel [33]. A parameter  $\alpha$  controls the sharpness of the sigmoid.

In addition, the **Modularity matrix**  $\mathbf{Q}$  is used as last baseline (Modularity). The matrix is computed by  $\mathbf{Q} = \mathbf{A} - \frac{\mathbf{d}\mathbf{d}^T}{\text{vol}}$  where  $\mathbf{d}$  contains the node degrees and the constant  $\text{vol}$  is the volume of the graph (see, e.g., [88]).

### Datasets

A collection of 22 datasets, representing labeled networks, is investigated for the experimental comparisons of the dissimilarity measures. The collection includes Zachary’s karate club [125], the

Dataset (labeled network)				Task	
Name	Labels	Nodes	Edges	Clustering	Classification
Dolphin_2	2	62	159	X	
Dolphin_4	4	62	159	X	
Football	12	115	613	X	
LFR1	3	600	6142	X	
LFR2	6	600	4807	X	
LFR3	6	600	5233	X	
IMDB	2	1126	20282		X
Newsgroup_2.1	2	400	33854	X	X
Newsgroup_2.2	2	398	21480	X	X
Newsgroup_2.3	2	399	36527	X	X
Newsgroup_3.1	3	600	70591	X	X
Newsgroup_3.2	3	598	68201	X	X
Newsgroup_3.3	3	595	64169	X	X
Newsgroup_5.1	5	998	176962	X	X
Newsgroup_5.2	5	999	164452	X	X
Newsgroup_5.3	5	997	155618	X	X
Political books	3	105	441	X	
WebKB-Cornell	6	346	13416		X
WebKB-Texas	6	334	16494		X
WebKB-Washington	6	434	15231		X
WebKB-Wisconsin	6	348	16625		X
Zachary	2	34	78	X	

**Table 1:** Datasets (networks) used for the experiments.

Dolphin datasets [77, 78], the Football dataset [42], the Political books<sup>13</sup>, three LFR benchmarks [68], the WebKB datasets [79], the IMDB dataset [79], and 9 Newsgroup datasets [69, 123].

The list of datasets along with their main characteristics is available in Table 1. Please note that all the datasets have not been used in both clustering and classification context. The two last columns indicate the investigated task for each dataset.

### 5.3 Node clustering experiment

We first describe the node clustering application together with the experimental methodology.

#### Evaluation metrics

Each partition will be assessed by comparing it with the observed “true partition” of the dataset. The following standard criteria will be used to evaluate the similarity between both partitions.

- ▶ The **Normalized Mutual Information (NMI)** [37, 104] between two partitions  $\mathcal{U}$  and  $\mathcal{V}$  is computed by dividing the mutual information [24] between the two partitions by the average of the respective entropy of  $\mathcal{U}$  and  $\mathcal{V}$ . See also [80].
- ▶ The **Adjusted Rand Index (ARI)** [54] is an extension of the Rand Index [95], which measures the degree of overlap between two partitions. The ARI has an expected value of 0, which is not the case for the initial Rand Index.

<sup>13</sup>Collected by V. Krebs and labelled by M.E. Newman, this dataset is not published, but available for download at <http://www-personal.umich.edu/~mejn/netdata/>.



Algorithm	Parameter values
FE RSP	$\theta = (0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 3, 5, 10, 15, 20)$
FETsallis RSPTsallis	$\theta = (10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5)$
LF	$\alpha = (0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 3, 5, 10, 15, 20)$
Katz	$\alpha = (0.05, 0.10, \dots, 0.95) \times (\rho(\mathbf{A}))^{-1}$
lCom	$t = (0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10)$
SCT	$\alpha = (5, 10, 15, \dots, 50)$

**Table 2:** Parameter range for the investigated methods.

## Experimental methodology

The experiment relies on a kernel  $k$ -means (see e.g., [34, 123]). For the dissimilarities, the followed methodology is similar to the one used in [102] (see this work for more details). More specifically, for each given dataset, the dissimilarity matrix  $\mathbf{D}$  obtained by the different methods is transformed into a kernel  $\mathbf{K}$  (a inner product matrix) using classical multidimensional scaling [11]. If the resulting kernel is not positive semi-definite, we simply set the negative eigenvalues to zero when computing the kernel. As a second step, a kernel  $k$ -means (see e.g., [34, 123]) is run 30 times on  $\mathbf{K}$ . The NMI and ARI are computed on the partition maximizing the modularity among these 30 trials. Recall that modularity is an unsupervised measure of the quality of a partition of the nodes (a set of communities) [88].

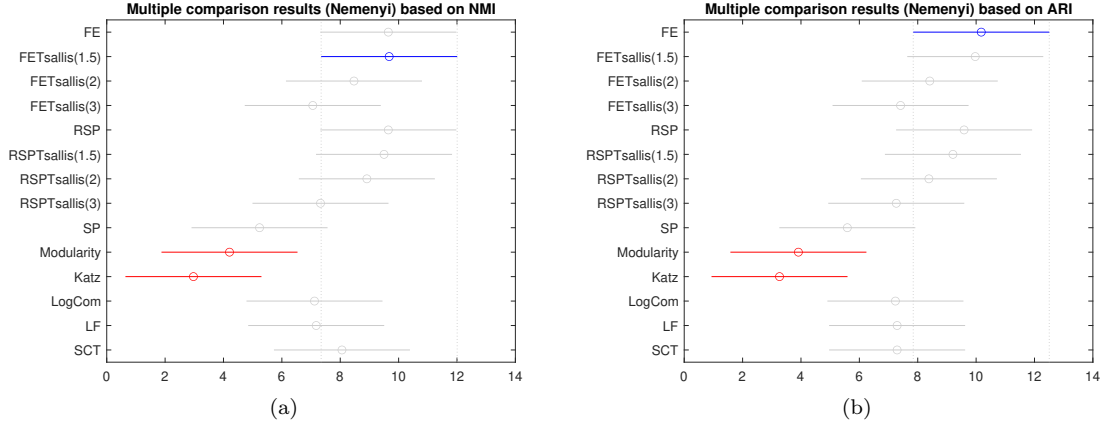
This operation is repeated 30 times (leading to a total of 900 runs of the  $k$ -means) to obtain the average modularity, NMI and ARI scores over these 30 repetitions for a given method (dissimilarity matrix), with a given value of its parameter (for instance,  $\theta$  in the case of methods based on RSP), on a specific dataset. Finally, the reported NMI and ARI score for each method and dataset is the average (over the 30 repetitions) for the parameter value leading to the largest modularity. Thus, modularity (which is unsupervised) is used as a metrics to tune the parameters of the algorithms [103]. The parameters that are tuned are the  $\theta$  for the FE and the RSP, in both standard (FE and RSP) and Tsallis versions (FETsallis and RSPTsallis), the  $\alpha$  for the LF, the  $\alpha$  for Katz, the  $t$  for the lCom, and finally the  $\alpha$  for the sigmoid transform of the SCT. The range of values that are tested strongly differs from the standard to the Tsallis version as we observed that the Tsallis version is less sensitive to variations in  $\theta$ . The values tested for these parameter are listed in the Table 2.

For the parameter  $r$  from the Tsallis regularization, three different values are tested  $r = \{1.5, 2, 3\}$ . Only a few values are investigated because the computation of the Tsallis-based dissimilarities is much slower that the one based on the KL divergence. Moreover, the results for these three values are displayed separately in order to analyse their impact on the results.

## Experimental results

The different methods are assessed globally across all datasets using the same method as in [102], based on a non-parametric Friedman-Nemenyi test [27]. The results for the ARI and the NMI are shown in Figure 3. In addition, a Wilcoxon signed-rank test [121] is performed pairwise to measure the significance (at level  $\alpha = 0.05$ ) of the differences observed in the algorithms' performance.

The figure shows that using a parameter  $r = 1.5$  for the Tsallis regularization tends to yield the best results out of the three values of  $r$  that were investigated. What concerns the FE distance, this



**Figure 3:** Clustering experiment. Mean ranks and 95% Nemenyi confidence intervals for the 11 methods across the 17 datasets, according to the NMI (a) and the ARI (b) performance measures. Two methods are considered as significantly different if their confidence intervals do not overlap. The best method is highlighted.

$r$  value allows to slightly outperform the FE relying on KL divergence based on the NMI, but it is not the case when considering the ARI. In both cases, the difference is not significant according to the Wilcoxon signed-rank test. More generally, except for the value  $r = 3$ , no significant difference can be observed, in terms of ranks, between the KL and the Tsallis divergence regularization for both the FE and the RSP from the pairwise Wilcoxon signed-rank tests.

The FETsallis with  $r = 1.5$  significantly outperforms the Modularity and the Katz kernel according to the Friedman-Nemenyi test. Additionally, according to the Wilcoxon, the difference in performance with the SP, the LogCom and the LF are significant as well.

Thus, what concerns the clustering task and the investigated datasets, the Tsallis regularization yields competitive results with respect to methods that have been shown to perform well in a context of kernel  $k$ -means clustering [102].

## 5.4 Semi-supervised classification experiment

We now turn to the semi-supervised classification experiment.

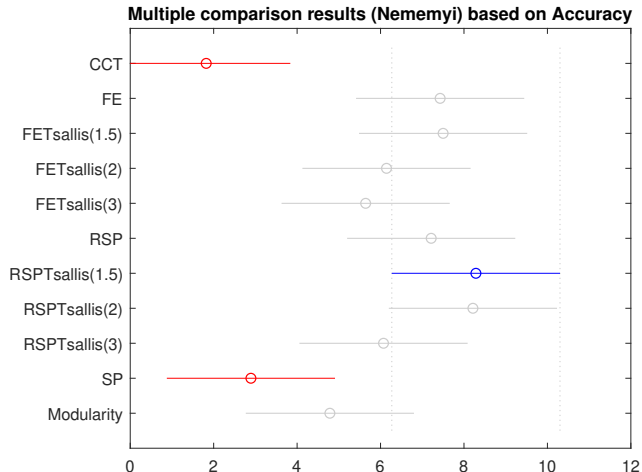
### Evaluation metrics

Each method will be evaluated in terms of classification accuracy on semi-supervised tasks where a subset of nodes of the graph is kept unlabeled (i.e. hidden). Then, the predicted labels of these unlabeled nodes are compared to the true, observed, labels which were hidden.

### Experimental methodology

We followed the same experimental methodology as in [36, 45]. More precisely, this graph-based semi-supervised classification methodology consists in extracting the five<sup>14</sup> dominant eigenvectors

<sup>14</sup>We arbitrarily report the results for 5 dimensions but also performed experiments with more dimensions with similar conclusions.



**Figure 4:** Semi-supervised experiment. Mean ranks and 95% Nemenyi confidence intervals for the 14 methods (see Table 3) across the 14 datasets. Two methods are considered as significantly different if their confidence intervals do not overlap. The best method overall (RSPTSallis(1.5)) is highlighted.

of a kernel matrix, derived from the dissimilarity matrix by classical multidimensional scaling, in order to use them as node features in a linear support vector machine (SVM). Note that this setting is inspired by the work of Zhang et al. [126, 127] as well as Tang et al. [105, 106, 107] who compute the dominant eigenvectors (a “latent social space”) of graph kernels or similarity matrices and then input them into a supervised classification method, such as a logistic regression or a SVM, to categorize the nodes.

All the methods are tested by using a standard  $5 \times 5$  nested cross-validation methodology. Each external cross-validation contains 5 folds, and methods are tested with a labelling rate of 20%. To tune parameters (see Table 2 for values), an internal 5-fold cross-validation on the training fold is performed with a labelling rate of 80%. The whole cross-validation procedure is repeated 5 times for different random permutations of the data, inducing different sets of labeled/unlabeled nodes. The *final accuracy* of the classifier on the investigated dataset is then obtained by averaging the results over the five repetitions, and is reported in Table 3.

Concerning the parameter  $r$  of the Tsallis regularization, as for clustering, we tested three values  $r = \{1.5, 2, 3\}$ . For the SVM, the margin parameter is tuned on the set of values  $c = \{10^{-2}, 10^{-1}, 1, 10, 100\}$ .

## Experimental results

The results of this experiment are reported in Table 3. The highest accuracy is highlighted in boldface for each dataset. As can be seen, the best method is dataset-dependent and no obvious, global, pattern is present. Therefore, in order to rate globally the results of each method, as before, we perform a nonparametric Friedman-Nemenyi statistical test and a Wilcoxon signed-rank tests at a level of confidence of 95% ( $\alpha = 0.05$ ) [27]. The results of the Nemenyi test are shown in Figure 4.

As in the clustering experiment, the figure shows that using a parameter  $r = 1.5$  for the Tsallis

Classif. method → Dataset ↓	FE	FETsallis(1.5)	FETsallis(2)	FETsallis(3)	RSP	RSPTsallis(1.5)	RSPTsallis(2)
IMDB	75.31	75.57	77.18	78.09	76.07	76.47	76.55
Newsgroup.2.1	96.79	96.04	95.44	95.06	95.96	95.75	95.56
Newsgroup.2.2	91.21	92.60	93.10	92.55	91.22	92.81	<b>93.13</b>
Newsgroup.2.3	95.98	96.07	96.30	96.15	95.39	96.18	96.35
Newsgroup.3.1	92.53	92.92	92.86	93.18	92.88	92.94	92.83
Newsgroup.3.2	93.45	93.19	92.83	92.65	93.34	<b>93.57</b>	93.07
Newsgroup.3.3	<b>93.66</b>	93.35	92.29	91.83	92.91	93.53	93.03
Newsgroup.5.1	<b>88.70</b>	88.54	88.24	87.62	88.09	88.22	88.03
Newsgroup.5.2	82.32	<b>82.55</b>	82.38	80.83	81.49	82.24	81.23
Newsgroup.5.3	80.27	82.69	82.78	77.53	80.38	82.65	<b>83.42</b>
WebKB-Cornell	57.20	57.37	56.52	54.44	57.53	58.80	58.31
WebKB-Texas	73.35	74.34	72.32	69.79	75.24	74.96	72.82
WebKB-Washington	68.15	66.53	66.26	63.85	68.88	68.40	66.15
WebKB-Wisconsin	74.33	70.69	68.58	68.66	73.91	71.67	69.04
Classif. method → Dataset ↓	RSPTsallis(3)	SP	Modularity	Katz	LogCom	LF	SCT
IMDB	<b>78.53</b>	74.68	74.37	68.93	76.28	73.87	78.36
Newsgroup.2.1	94.51	93.58	95.85	95.15	96.18	96.65	<b>97.14</b>
Newsgroup.2.2	92.79	90.36	91.22	90.38	91.18	90.36	91.14
Newsgroup.2.3	96.32	<b>96.78</b>	95.78	93.21	95.43	95.54	95.79
Newsgroup.3.1	92.54	93.01	93.02	91.00	<b>94.00</b>	92.98	93.99
Newsgroup.3.2	92.85	89.32	92.63	91.25	92.17	92.15	92.35
Newsgroup.3.3	92.52	91.14	91.20	88.73	91.49	90.93	93.35
Newsgroup.5.1	87.59	86.60	77.04	79.74	86.30	86.22	87.29
Newsgroup.5.2	80.27	78.36	75.97	64.47	79.04	80.10	80.45
Newsgroup.5.3	82.31	72.73	76.51	66.33	78.65	79.94	80.38
WebKB-Cornell	55.62	47.36	50.71	52.33	<b>59.16</b>	58.35	57.30
WebKB-Texas	72.93	61.20	73.01	67.89	<b>75.45</b>	74.18	74.27
WebKB-Washington	64.69	52.35	62.52	64.56	<b>69.92</b>	67.47	67.66
WebKB-Wisconsin	69.64	62.46	73.42	73.61	<b>74.99</b>	74.48	72.77

**Table 3:** Classification accuracies in percent for the various classification methods obtained on the different datasets. For each dataset and method, the final accuracy is obtained by averaging over 5 repetitions of a standard cross-validation procedure. Each repetition consists of a nested cross-validation with 5 external folds (test sets, for validation) on which the accuracy of the classifier is averaged, and 5 internal folds (for parameter tuning). The best performing method is highlighted in boldface for each dataset.

regularization tends to yield the best results out of the three values of  $r$  that were investigated. Furthermore, the Wilcoxon signed-rank tests show that the FETsallis and the RSPTsallis perform significantly better than the SP, the Katz and the Modularity (except for the FETsallis with a parameter  $r = 3$ ). The tests also show that the RSPTsallis with a parameter  $r = 1.5$  obtains significantly better results than the LF. Nevertheless, except for the FETsallis with a parameter  $r = 3$ , no significant difference can be observed, in terms of ranks, between the KL and the Tsallis divergence regularization for both the FE and the RSP.

This confirms that the Tsallis RSP and FE dissimilarity measures, and especially the RSP, are able to capture the community structure of the graph in an accurate way, at least on the investigated datasets.

## 6 Conclusion and future work

This paper showed that sparse randomized routing policies in a network can be obtained when regularizing the least-cost routing by the Tsallis divergence instead of the Kullback-Leibler divergence. Two different algorithms are detailed, a simple and faster procedure for the case  $r = 2$  based on a linear search and a slower one for the more general case  $r > 1$  based on a bisection search technique. Indeed, in practice, we observed that the bisection method is significantly slower than the linear search method, at least on the investigated datasets.

In that context, various interesting quantities can be derived from the routing policy, especially the expected cost and the minimized free energy between the source and the target nodes. These quantities can be used as dissimilarity measures between the nodes of the network for tackling pattern recognition and machine learning tasks. A nice property is that they interpolate between the shortest path distance (when  $\theta \rightarrow \infty$ ) and the commute-cost distance (proportional to the resistance distance,  $\theta \rightarrow 0^+$ ). Another interesting property is the fact that, as the standard randomized shortest paths, they are taking the degree of inter-connectivity (direct and indirect), in addition to proximity, into account in the computation of the dissimilarity.

Indeed, it is well-known that the standard shortest path distance and the resistance distance [64], while very useful in many contexts, show important drawbacks in some situations which hinders their use as distance measures between nodes in some applications. More precisely, the shortest path distance does not integrate the concept of high connectivity between the two nodes (it only considers the shortest paths, see, e.g., [34]), while the resistance distance provides useless results when dealing with large graphs (the so-called “lost-in-space effect” [117, 118]). Another drawback of the shortest path distance is that it usually provides a large amount of ties when comparing distances, especially on unweighted and undirected graphs. Moreover, it has been shown recently [47] that the FE distance based on Kullback-Leibler regularization (called the logarithmic Laplace transformed hitting time in their paper) avoids to a certain extend the “lost-in-space effect”. Therefore, we conjecture that the FE dissimilarity based on the Tsallis divergence introduced in this paper benefits from the same property.

Experimental comparisons based on two pattern recognition tasks show that the proposed distances are competitive with other state-of-the-art techniques. The main drawback, however, is the fact that the computation of the dissimilarities in the general  $r \neq 2$  case is time-consuming, preventing its application on large graphs.

Further work will be devoted to the improvement of the algorithm computing the FE distance based on the Tsallis divergence in the  $r \neq 2$  case. For instance, we could adopt a mixed strategy by first applying the line search in order to identify the one-unit integer interval in which the optimal value lies, and then running a bisection search within this interval. Another idea would be to use the algorithm proposed in [30] for computing the orthogonal projection on the unit simplex, based on a modification of the randomized median finding procedure [10, 23]. The link between the orthogonal projection on the unit simplex and our formulation (12) should also be studied.

Still another interesting contribution would be to apply the Tsallis divergence regularization for solving Markov decision problems, inducing sparse policies; therefore extending the work of [71]. We also plan to use the Tsallis divergence for the design of algorithms solving the sparse randomized optimal transport on a graph problem, extending previous work [46]. Indeed, in the same way as in this paper, the Kullback-Leibler divergence can be replaced by the Tsallis divergence as regularization term in the model, therefore providing sparse routing policies.

## Acknowledgements

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, Belgium. We thank these institutions for giving us the opportunity to conduct both fundamental and applied research. We also thank Professor Masashi Shimbo and Dr Amin Mantrach for the helpful references and discussions.

---

## Appendices

These appendices discuss the alternative form of the randomized shortest paths, the convexity of the objective function with Tsallis divergence regularization, as well as the algorithms for computing the spmin function appearing in Equation (13), returning transition probabilities  $\mathbf{p}_i$  associated to a node  $i$ . To the best of our knowledge, these algorithms were first studied in [57, 85] (although we suspect that they have probably been investigated before). This appendix provides a reformulation of the relevant material contained in these papers (C.1 – C.2), as well as an extension of their algorithms for dealing with an arbitrary reference distribution (and thus regularizing with Tsallis divergence instead of Tsallis entropy, C.3).

### A An alternative view of the standard randomized shortest path framework

The path-based formalism (2) can be transformed into a “local” form (see [3, 39, 99]) which will be used for deriving the sparse RSP. In this new form, the policy is computed in an iterative way by exploiting Lagrange duality.

#### A.1 Alternative form of the objective function

To this end, let us introduce  $\eta((i, j) \in \varphi)$  defined as the number of times edge  $(i, j)$  is visited along path  $\varphi$ . Because the probability of a path can be expressed as a product of transition probabilities (see the discussion after Equation (4)), the path-based quantities  $\log(P(\varphi)/\tilde{\pi}(\varphi))$  and  $\tilde{c}(\varphi)$  can be expressed as

$$\begin{cases} \log \frac{P(\varphi)}{\tilde{\pi}(\varphi)} = \sum_{(i,j) \in \mathcal{E}} \eta((i, j) \in \varphi) \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} \\ \tilde{c}(\varphi) = \sum_{(i,j) \in \mathcal{E}} \eta((i, j) \in \varphi) c_{ij} \end{cases} \quad (21)$$

Injecting these relations in the objective function appearing in Equation (2) yields

$$\begin{aligned} \phi_{st}^{\text{KL}}(\mathbf{P}) &= \sum_{\varphi \in \mathcal{P}_{st}} P(\varphi) \tilde{c}(\varphi) + T \sum_{\varphi \in \mathcal{P}_{st}} P(\varphi) \log \left( \frac{P(\varphi)}{\tilde{\pi}(\varphi)} \right) \\ &= \sum_{\varphi \in \mathcal{P}_{st}} P(\varphi) \sum_{(i,j) \in \mathcal{E}} \eta((i, j) \in \varphi) c_{ij} + T \sum_{\varphi \in \mathcal{P}_{st}} P(\varphi) \sum_{(i,j) \in \mathcal{E}} \eta((i, j) \in \varphi) \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} \\ &= \sum_{(i,j) \in \mathcal{E}} \bar{n}_{ij} c_{ij} + T \sum_{(i,j) \in \mathcal{E}} \bar{n}_{ij} \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} \\ &= \sum_{(i,j) \in \mathcal{E}} \bar{n}_{ij} p_{ij} \left( c_{ij} + T \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right) \end{aligned} \quad (22)$$

where  $\bar{n}_{ij} \triangleq \sum_{\varphi \in \mathcal{P}_{st}} P(\varphi) \eta((i, j) \in \varphi)$  is the expected number of passages (the flow) through edge  $(i, j)$  and we used  $\bar{n}_{ij} = \bar{n}_i p_{ij}$  with  $\bar{n}_i = \bar{n}_{i\bullet} = \sum_{j \in \text{Succ}(i)} \bar{n}_{ij}$  denoting the expected number of visits to node  $i$ . This comes from the fact that the expected number of passages through edge  $(i, j)$  is equal to the number of visits to node  $i$  times the probability of following the link  $(i, j)$  from node  $i$ . This formulation of the RSP problem is also closely related to the framework of [7, 44] based on edge flows and flow conservation.

The objective function (22) should be minimized with respect to the (local) policy, that is, the set of transition probabilities  $p_{ij}$  associated to edges. It is shown in [3] that this objective function is strictly convex with respect to edge flows,  $\bar{n}_{ij} = \bar{n}_i p_{ij}$ . Indeed, the objective function, which is expressed in (22) in terms of transition probabilities and number of visits to nodes, can also be expressed in function of edge flows only, or in function of transition probabilities only. Therefore, because the correspondence between edge flows and transition probabilities is differentiable and one-to-one for a unit input flow (which is indeed the case in the RSP model), any stationary point of (22) with respect to the transition probabilities  $p_{ij}$  is also a stationary point with respect to the corresponding edge flows. Thus, because the objective function is convex with respect to the  $\bar{n}_{ij}$  and the domain is convex, it must be a global minimum.

Interestingly, this also shows that the path-based formalism of Equation (2) is equivalent to minimizing the local cost plus KL divergence,  $c_{ij} + T \log(p_{ij}/p_{ij}^{\text{ref}})$ , which is also the purpose of Kullback-Leibler, or path integral, control developed in the field of reinforcement learning and control theory. Therefore, as already mentioned in the related work (Subsection 1.2), the randomized shortest paths framework is equivalent to some of these models developed in reinforcement learning (see, e.g., [15, 35, 59, 98, 111, 112]), as initiated by [113, 114].

## A.2 Computing the optimal policy

In this subsection, an algorithm for computing the optimal transition probabilities (the policy) is developed. It aims at minimizing the objective function (22) by considering the transition probabilities and the expected number of visits as independent. The dependency between the two quantities is introduced as a constraint in the formulation, as commonly done in discrete-state discrete-time optimal control (see, e.g., [75]). After *renumbering the nodes* in such a way that node 1 is the source node and node  $n$  the target node<sup>15</sup> for convenience, this leads to the following Lagrange function only including the equality constraints

$$\begin{aligned} \mathcal{L}(\mathbf{P}, \bar{\mathbf{n}}; \boldsymbol{\mu}, \boldsymbol{\lambda}_{\text{KL}}) &= \sum_{i \in \mathcal{V} \setminus n} \bar{n}_i \sum_{j \in \text{Succ}(i)} p_{ij} \left( c_{ij} + T \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right) \\ &+ \sum_{i \in \mathcal{V} \setminus n} \mu_i \left( 1 - \sum_{j \in \text{Succ}(i)} p_{ij} \right) \\ &+ \sum_{j \in \mathcal{V}} \lambda_j^{\text{KL}} \left( \sum_{i \in \text{Pred}(j)} \bar{n}_i p_{ij} + \delta_{1j} - \bar{n}_j \right) \end{aligned} \quad (23)$$

where  $\text{Succ}(i)$  is the set of successor nodes<sup>16</sup> of node  $i$  and  $\text{Pred}(j)$  is the set of predecessor nodes of node  $j$ . The quantities  $\mu_i, \lambda_i^{\text{KL}}$  are standard Lagrange parameters dealing with equality

<sup>15</sup>It is assumed that the source node is different from the target node.

<sup>16</sup>Recall that target node  $n$  is killing and absorbing, and thus has no successor. It is therefore not the predecessor of any node.

constraints. The transition probabilities and the expected number of visits are therefore considered as independent in the optimization of the Lagrange function and must be non-negative. The relation between  $\bar{n}_i$  and  $p_{ij}$  is given by the system of linear equations computing the expected number of visits to nodes in an absorbing Markov chain,  $\bar{n}_j = \sum_{i \in \mathcal{P}red(j)} \bar{n}_i p_{ij} + \delta_{1j}$  for each  $j \in \mathcal{V}$ , when a unit flow is injected in node 1 (see, e.g., [90, 108]). By the property of flow conservation in a Markov chain, it is clear that  $\bar{n}_n = 1$  for the target node.

Our procedure optimizes sequentially the objective function by Lagrange duality as follows [8, 25, 43, 83]. We first minimize the Lagrange function with respect to the transition probabilities  $p_{ij}$  subject to sum-to-one constraints, while fixing the Lagrange parameters  $\lambda_i^{\text{KL}}$ . We will observe that they only depend on the Lagrange parameters. Lagrange parameters are then computed by maximizing the dual, a common optimization procedure called the Arrow-Hurwicz-Uzawa algorithm [5]. The two steps are iterated until convergence, which is guaranteed because each sub-problem reaches its optimum uniquely [9]. In practice, we observed that the duality gap is always zero, showing that a global minimum is reached.

### Computation of the transition probabilities

For the estimation of the transition probabilities step, there is no need to introduce non-negativity constraints because KL divergence regularization ensures that the estimates satisfy the constraint [60]. Taking the partial derivative of the Lagrange function (23) with respect to the transition probabilities associated to edges and setting the result to zero provides

$$T \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} = \frac{\mu_i}{\bar{n}_i} - T - (c_{ij} + \lambda_j^{\text{KL}})$$

Then, using  $\theta = 1/T$ , isolating the transition probabilities as well as imposing the sum-to-one constraint yields

$$p_{ij} = \frac{p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \lambda_j^{\text{KL}})]}{\sum_{k \in \text{Succ}(i)} p_{ik}^{\text{ref}} \exp[-\theta(c_{ik} + \lambda_k^{\text{KL}})]} \quad \text{for all } (i, j) \in \mathcal{E} \quad (24)$$

which only depends on the Lagrange parameters  $\lambda_i^{\text{KL}}$ . This corresponds to the “local” optimal randomized policy for going from 1 to  $n$ , according to KL divergence regularization. Let us now compute these Lagrange parameters.

### Computation of the Lagrange parameters

Computing the  $\lambda_i^{\text{KL}}$  aims at solving the dual problem. Indeed, by defining respectively the expected cost and the KL divergence per node,  $\tilde{c}_i = \sum_{j \in \text{Succ}(i)} p_{ij} c_{ij}$  and  $\bar{h}_i^{\text{KL}} = \sum_{j \in \text{Succ}(i)} p_{ij} \log(p_{ij}/p_{ij}^{\text{ref}})$  for  $i \neq n$ , together with  $\tilde{c}_n = 0$  and  $\bar{h}_n^{\text{KL}} = 0$  for target node, the problem of computing the expected number of visits  $\bar{n}_i$  when transition probabilities are fixed can be reformulated from (23) as a linear programming problem: minimize  $(\tilde{\mathbf{c}} + T\bar{\mathbf{h}}_{\text{KL}})^T \bar{\mathbf{n}}$  with respect to  $\bar{\mathbf{n}}$  subject to the constraints  $(\mathbf{I} - \mathbf{P})^T \bar{\mathbf{n}} = \mathbf{e}_1$  and  $\bar{\mathbf{n}} \geq \mathbf{0}$ .

However, because we instead need the vector of Lagrange parameters  $\lambda_{\text{KL}}$  in order to compute the transition probabilities (see (24)), we are more interested in the dual problem (see, e.g., [43],



page 184<sup>17</sup>),

$$\left\{ \begin{array}{l} \text{maximize} \quad \mathbf{e}_1^T \boldsymbol{\lambda}_{\text{KL}} \\ \text{subject to} \quad (\mathbf{I} - \mathbf{P}) \boldsymbol{\lambda}_{\text{KL}} = \bar{\mathbf{c}} + T \bar{\mathbf{h}}_{\text{KL}} \\ \boldsymbol{\lambda}_{\text{KL}} \geq \mathbf{0} \end{array} \right. \quad (25)$$

where  $\mathbf{e}_1$  is a column vector full of 0's, except in position 1 containing a 1. Because the elements of  $(\mathbf{I} - \mathbf{P})^{-1} = \mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \dots$  are all non-negative, the non-negativity constraint on the Lagrange parameters is automatically satisfied. These Lagrange parameters are thus obtained by solving the system of linear equations

$$(\mathbf{I} - \mathbf{P}) \boldsymbol{\lambda}_{\text{KL}} = \bar{\mathbf{c}} + T \bar{\mathbf{h}}_{\text{KL}} \quad (26)$$

where, in matrix form,  $\bar{\mathbf{c}} = (\mathbf{P} \circ \mathbf{C}) \mathbf{e}$  and  $\bar{\mathbf{h}}_{\text{KL}} = (\mathbf{P} \circ (\log \mathbf{P} - \log \mathbf{P}_{\text{ref}})) \mathbf{e}$ , with  $\mathbf{e}$  being a column vector of 1's and  $\circ$  the elementwise matrix product. Elementwise, we have

$$\lambda_i^{\text{KL}} = \sum_{j \in \text{Succ}(i)} p_{ij} \left( c_{ij} + \lambda_j^{\text{KL}} + T \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right) \quad (27)$$

and notice that this implies  $\lambda_n^{\text{KL}} = 0$  for the target node  $n$ . The procedure aims at iterating Equations (24) and (26) until convergence. Interestingly, the Lagrange parameters have a nice interpretation, as explained in the next subsection.

### Interpretation of the Lagrange parameters

Let us now give an interpretation to the Lagrange parameters  $\boldsymbol{\lambda}_{\text{KL}}$ . From Equation (24), we directly obtain

$$T \log \frac{p_{ij}}{p_{ij}^{\text{ref}}} = -(c_{ij} + \lambda_j^{\text{KL}}) - T \log \sum_{k \in \text{Succ}(i)} p_{ik}^{\text{ref}} \exp[-\theta(c_{ik} + \lambda_k^{\text{KL}})]$$

By injecting this expression in Equation (27), we obtain

$$\lambda_i^{\text{KL}} = -\frac{1}{\theta} \log \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \lambda_j^{\text{KL}})] \quad (28)$$

which is nothing else than the Bellman-Ford-like recurrence formula<sup>18</sup> for computing the directed *free energy distance*  $\phi^{\text{KL}}$  (see<sup>19</sup> [36], Eq. (34)); therefore  $\boldsymbol{\lambda}^{\text{KL}} = \phi^{\text{KL}}$  and the Lagrange parameters are equal to the directed FE distances at the optimum.

This directed FE distance is the minimum free energy obtained by replacing path probabilities by the optimal ones (see Equation (3)) in the free energy objective function provided by Equation (2) (see [63] for details). This quantity has many interesting properties: (1) it plays the role of a potential at the continuous space-time limit [40], (2) it defines a distance measure between nodes when symmetrized [36, 63], (3) it interpolates between the least-cost and the commute-cost distance, (4) it can be interpreted as minus  $T$  times the log-likelihood of surviving during a particular killed random walk [36], and (5) it performed consistently well in a number of pattern recognition tasks [36, 102, 103], among others.

<sup>17</sup>Note that the equation in [43] also holds for equality constraints, providing  $(\mathbf{I} - \mathbf{P})^T \bar{\mathbf{n}} = \mathbf{e}_1$ , which is the case here.

<sup>18</sup>The min operator in the Bellman-Ford expression is replaced by a softmin operator, see [36].

<sup>19</sup>As it plays the role of a potential, the FE distance was called the potential distance in [36].

## B About the convexity of the Tsallis-regularized objective function

This section discusses the convexity of the Tsallis-regularized free energy objective function appearing in Equation (9). To this end, we will compute the Hessian matrix and verify if the corresponding quadratic form is positive semi-definite.

Using  $\bar{n}_{ij} = \bar{n}_i p_{ij}$  with  $\bar{n}_i = \bar{n}_{i\bullet} = \sum_{j \in \text{Succ}(i)} \bar{n}_{ij}$ , we first reformulate this objective function in terms of edge flows  $\bar{n}_{ij}$ ,

$$\begin{aligned} \phi_{st}^{\text{Ts}} &= \sum_{i \in \mathcal{V}} \bar{n}_{i\bullet} \sum_{j \in \text{Succ}(i)} p_{ij} \left( c_{ij} + \frac{T}{r-1} \left( \left( \frac{p_{ij}}{p_{ij}^{\text{ref}}} \right)^{r-1} - 1 \right) \right) \\ &= \sum_{i \in \mathcal{V}} \sum_{j \in \text{Succ}(i)} \left( c_{ij} \bar{n}_{ij} + \frac{T}{r-1} \frac{(\bar{n}_{ij})^r}{(p_{ij}^{\text{ref}} \bar{n}_{i\bullet})^{r-1}} - \frac{T}{r-1} \bar{n}_{ij} \right) \end{aligned} \quad (29)$$

For computing the Hessian, only the central term is meaningful. We will therefore study the function (at first, we do not consider the reference probability  $p_{ij}^{\text{ref}}$  for simplicity),

$$f(\mathbf{N}) = \sum_{k \in \mathcal{V}} \sum_{l \in \text{Succ}(k)} \frac{(\bar{n}_{kl})^r}{(\bar{n}_{k\bullet})^{r-1}} \quad (30)$$

The first-order partial derivatives are

$$\frac{\partial f}{\partial \bar{n}_{ij}} = r p_{ij}^{r-1} - (r-1) \sum_{j' \in \text{Succ}(i)} p_{ij'}^r \quad (31)$$

where  $p_{ij}$  denotes  $\bar{n}_{ij}/\bar{n}_{i\bullet}$  for simplicity. Then the Hessian is

$$h_{(i,j)(k,l)} = \frac{\partial^2 f}{\partial \bar{n}_{kl} \partial \bar{n}_{ij}} = r(r-1) \frac{\delta_{ik}}{\bar{n}_{i\bullet}} \left( \delta_{jl} p_{ij}^{r-2} - (p_{ij}^{r-1} + p_{il}^{r-1}) + \sum_{j' \in \text{Succ}(i)} p_{ij'}^r \right) \quad (32)$$

Thus, the following quadratic form should be non-negative for any value of  $x_{(i,j)}$ ,

$$\begin{aligned} Q &= \sum_{(i,j) \in \mathcal{E}} \sum_{(k,l) \in \mathcal{E}} x_{(i,j)} h_{(i,j)(k,l)} x_{(k,l)} \\ &= r(r-1) \sum_{i \in \mathcal{V}} \frac{1}{\bar{n}_{i\bullet}} \sum_{j,l \in \text{Succ}(i)} x_{(i,j)} \underbrace{\left( \delta_{jl} p_{ij}^{r-2} - (p_{ij}^{r-1} + p_{il}^{r-1}) + \sum_{j' \in \text{Succ}(i)} p_{ij'}^r \right)}_{\text{matrix } q_{jl}(i)} x_{(i,l)} \end{aligned} \quad (33)$$

and therefore it suffices to show that the matrices  $\mathbf{Q}(i) = (q_{jl}(i))$  are positive semi-definite in order to prove convexity ( $\bar{n}_{i\bullet}$  is always positive). When introducing the reference probabilities, the objective function becomes

$$f(\mathbf{N}) = \sum_{k \in \mathcal{V}} \sum_{l \in \text{Succ}(k)} (p_{kl}^{\text{ref}})^{1-r} \frac{(\bar{n}_{kl})^r}{(\bar{n}_{k\bullet})^{r-1}} \quad (34)$$

and the Hessian is

$$\begin{aligned}
h_{(i,j)(k,l)} &= r(r-1) \frac{\delta_{ik}}{\bar{n}_{i\bullet}} \left( \delta_{jl} (p_{ij}^{\text{ref}})^{1-r} p_{ij}^{r-2} - ((p_{ij}^{\text{ref}})^{1-r} p_{ij}^{r-1} + (p_{il}^{\text{ref}})^{1-r} p_{il}^{r-1}) \right. \\
&\quad \left. + \sum_{j' \in \text{Succ}(i)} (p_{ij'}^{\text{ref}})^{1-r} p_{ij'}^r \right)
\end{aligned} \tag{35}$$

Then, the quadratic form can be readily deduced

$$\begin{aligned}
Q &= r(r-1) \sum_{i \in \mathcal{V}} \frac{1}{\bar{n}_{i\bullet}} \sum_{j,l \in \text{Succ}(i)} x_{(i,j)} \left( \delta_{jl} (p_{ij}^{\text{ref}})^{1-r} p_{ij}^{r-2} - ((p_{ij}^{\text{ref}})^{1-r} p_{ij}^{r-1} + (p_{il}^{\text{ref}})^{1-r} p_{il}^{r-1}) \right. \\
&\quad \left. + \sum_{j' \in \text{Succ}(i)} (p_{ij'}^{\text{ref}})^{1-r} p_{ij'}^r \right) x_{(i,l)}
\end{aligned} \tag{36}$$

Because we did not find an easy way to prove formally the positive semi-definiteness of the matrices  $\mathbf{Q}(i)$  with elements  $q_{jl}(i) = \delta_{jl} (p_{ij}^{\text{ref}})^{1-r} p_{ij}^{r-2} - ((p_{ij}^{\text{ref}})^{1-r} p_{ij}^{r-1} + (p_{il}^{\text{ref}})^{1-r} p_{il}^{r-1}) + \sum_{j' \in \text{Succ}(i)} (p_{ij'}^{\text{ref}})^{1-r} p_{ij'}^r$ , we decided to test it numerically. More precisely, we generated  $10^6$  random instances of 20-dimensional probability vectors  $\mathbf{p}_i$  and  $\mathbf{p}_i^{\text{ref}}$ , as well as values of the  $r$  parameter in [1.1, 4.1]. The smallest eigenvalue of the corresponding  $\mathbf{Q}(i)$  matrix is then extracted. In all cases, the smallest eigenvalue was equal to  $\lambda_{\min} = 0$  (no negative eigenvalue), up to small errors of the order  $|\lambda_{\min}/\lambda_{\max}| < 10^{-14}$ . This provides evidence that the objective function is convex, although it remains a conjecture at this point.

## C Minimizing expected cost plus Tsallis free energy

In this section, we solve the problem stated in Equation (12), that is, the minimization of an expected cost with Tsallis divergence regularization. We proceed gradually in three steps. First, the  $r = 2$  case and a uniform reference probability distribution is considered (C.1). Then, we extend the results to the more general  $r > 1$  case (C.2). Finally, the most general case of  $r > 1$  and a non-uniform reference probability distribution (and thus Tsallis divergence regularization) of Equation (12) is considered (C.3). Note that C.1-C.2 are based on [57, 85].

### C.1 Results for the $r = 2$ case and a uniform reference probabilities

This subsection derives the algorithm for finding a sparse solution to the problem of minimizing an expected cost under quadratic constraints, thus in the case where  $r = 2$  and a uniform reference distribution (adapted from [57, 85]). The problem stated in Equation (12) then reduces to

$$\begin{cases} \text{minimize} & \mathbf{c}^T \mathbf{p} + T \|\mathbf{p}\|_2^2 \\ & \mathbf{p} \\ \text{subject to} & \mathbf{e}^T \mathbf{p} = 1 \\ & \mathbf{p} \geq \mathbf{0} \end{cases} \tag{37}$$

where  $T$  is the temperature parameter, and it is assumed that vectors  $\mathbf{p}$  and  $\mathbf{c} \geq \mathbf{0}$  contain  $m$  elements (corresponding to successor nodes of node  $i$ ). We omit the row index  $i$  as well as the quote for the augmented cost for the sake of simplicity. In our context, the vector  $\mathbf{c}$  corresponds to the augmented costs,  $\mathbf{c}'_i$ , associated to the outgoing edges of a transient node  $i$  and  $\mathbf{p}$  corresponds to its transition probabilities  $\mathbf{p}_i$  (see Equation (12)).

### The Karush-Kuhn-Tucker conditions

By using the necessary Karush-Kuhn-Tucker conditions [43, 76, 96] and introducing the Lagrange parameters  $\mu$  (equality constraint in (37)) and  $\boldsymbol{\lambda}$  (inequality constraints in (37)), we obtain, in addition to the sum-to-one and non-negativity constraints on  $\mathbf{p}$ ,

$$\begin{cases} \mathbf{c} + 2T\mathbf{p} - \mu\mathbf{e} - \boldsymbol{\lambda} = \mathbf{0} \\ \boldsymbol{\lambda}^\top \mathbf{p} = 0 \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \quad (38)$$

from which we immediately deduce  $\lambda_i p_i = 0$  for each  $i$ . In other words,  $\lambda_i = 0$  or  $p_i = 0$  because both  $\boldsymbol{\lambda}$  and  $\mathbf{p}$  are non-negative. We further denote by  $\mathcal{Q}_+^*$  the set of *strictly positive*  $p_i$  (to be found) and  $|\mathcal{Q}_+^*|$  the number of such elements.

Let us now consider the different cases. First, if  $p_i = 0$  for element  $i$ , Equation (38) tells us that  $\lambda_i = c_i - \mu$ . And because  $\lambda_i \geq 0$ , we must have  $c_i \geq \mu$  when  $p_i = 0$ . By taking the contraposition of the previous implication and using the fact that  $p_i \geq 0$ , we obtain that if  $c_i < \mu$  then  $p_i > 0$  (and thus also  $\lambda_i = 0$ ).

Next, we investigate the situation where  $c_i \geq \mu$ , equivalent to  $\mu - c_i \leq 0$ . From (38), this implies  $2Tp_i - \lambda_i \leq 0$  and thus  $p_i \leq \lambda_i/2T$ . Because the  $p_i$  are non-negative and  $\lambda_i = 0$  or  $p_i = 0$ , this implies that  $p_i = 0$  when  $c_i \geq \mu$ . The parameter  $\mu$  is therefore a *threshold* telling us when  $p_i$  should be put to zero.

Finally if, for element  $i$ ,  $p_i > 0$  then  $\lambda_i = 0$  and the non-negativity constraint on  $p_i$  is non-active. Then, from (38),  $c_i + 2Tp_i - \mu = 0$ , which provides  $p_i = \frac{1}{2T}(\mu - c_i)$ . Therefore, the  $p_i$  are of the following form

$$p_i = \frac{1}{2T}[\mu - c_i]_+ = \begin{cases} \frac{1}{2T}(\mu - c_i) & \text{when } \mu - c_i > 0 \\ 0 & \text{when } \mu - c_i \leq 0 \end{cases} \quad (39)$$

where  $[x]_+ = \max(x, 0)$ : if  $x$  is negative, it is replaced by 0. We observe that  $p_i$  is put to zero when  $(\mu - c_i)$  becomes non-positive. Therefore, without loss of generality, from now we assume a numbering such that the elements of the cost vector  $\mathbf{c}$  are *sorted and indexed by increasing value* of  $c_i$ ,  $c_1 \leq c_2 \leq \dots \leq c_m$ . In that case, from (39), the sequence of  $(p_i)_{i=1}^m$  is monotonic non-increasing for a fixed  $\mu$ .

The result (39) also tells us that the optimal set of strictly positive  $p_i$  is given by  $\mathcal{Q}_+^* = \{1, 2, \dots, k^*\}$ , for some index  $k^* = |\mathcal{Q}_+^*|$  to be found. We now have to compute the Lagrange parameter  $\mu$  as well as this threshold index  $k^*$ .

### Computing the Lagrange parameter $\mu$

By expressing the sum-to-one constraint on  $\mathbf{p}$  and assuming that the optimal number of strictly positive elements  $|\mathcal{Q}_+^*| = k^*$  is known, the Lagrange parameter can easily be computed,  $\mu = \frac{1}{|\mathcal{Q}_+^*|} \left( \sum_{j \in \mathcal{Q}_+^*} c_j \right) + \frac{2T}{|\mathcal{Q}_+^*|}$ . By injecting this expression into Equation (39), we obtain for  $i \in \mathcal{Q}_+^*$  and thus  $i \leq k^*$

$$p_i = \frac{1}{2T} \left( \frac{1}{|\mathcal{Q}_+^*|} \sum_{j \in \mathcal{Q}_+^*} c_j - c_i \right) + \frac{1}{|\mathcal{Q}_+^*|} = \frac{1}{2T} \left( \underbrace{\frac{1}{k^*} \sum_{j=1}^{k^*} c_j - c_i}_{=s(k^*)} \right) + \frac{1}{k^*} \quad (40)$$

and this expression can be extended to the whole set of  $p_i$ ,  $i = 1, \dots, m$ , with

$$p_i = \left[ \frac{1}{2T} \left( \frac{1}{k^*} \sum_{j=1}^{k^*} c_j - c_i \right) + \frac{1}{k^*} \right]_+ \quad (41)$$

where  $[x]_+ = \max(x, 0)$ . When  $T$  is large, we obtain a uniform distribution,  $p_i = 1/k^*$  for  $i = 1, \dots, k^*$ , whereas when  $T$  is close to zero, the probability mass is concentrated on the first element (the lowest cost),  $p_i = \delta_{i1}$ , a Kronecker delta. Let us now compute the threshold index  $k^*$ .

### Computing the threshold index $k^*$

The idea now for finding  $k^*$  is to investigate sequentially a number of positive elements  $k = 1, 2, \dots$  for  $\mathbf{p}$  and select the  $k$  for which the L1 norm of vector  $\mathbf{p}$  will be equal or greater than 1. More precisely, from Equation (39), assuming a number of positive elements  $k \leq k^*$  is selected, the sum of the corresponding entries of the vector  $\mathbf{p}$  (the L1 norm) is equal to

$$\sum_{i=1}^k p_i = \frac{1}{2T} \sum_{i=1}^k (\mu - c_i) = \frac{1}{2T} \left( k\mu - \sum_{i=1}^k c_i \right) \quad (42)$$

The optimal threshold  $\mu$  is found when the L1 norm is *exactly equal* to 1 (the only solution which is admissible). The idea is thus to increase  $k$  until<sup>20</sup> the L1 norm is equal to or exceeds one, as proposed in [57].

To this end, let us introduce the following auxiliary function from Equation (42), corresponding to the L1 norm when considering an increasing sequence of  $k = 1, 2, \dots$  and corresponding discrete values  $\mu = c_k$  for  $\mu$ ,

$$L_1(k) = \frac{1}{2T} \left( kc_k - \underbrace{\sum_{i=1}^k c_i}_{s(k)} \right) = \frac{1}{2T} (kc_k - s(k)) \quad (43)$$

with  $s(k) = \sum_{i=1}^k c_i$ . Obviously,  $L_1(1) = 0$ . Moreover, the function  $L_1$  is a monotonic non-decreasing function. Indeed,  $2T L_1(k+1) = (k+1)c_{k+1} - s(k+1) = kc_{k+1} + c_{k+1} - s(k) - c_{k+1} = kc_{k+1} - s(k) \geq kc_k - s(k) = 2T L_1(k)$ .

The successive terms are computed incrementally until  $L_1(k^*) < 1$  and  $L_1(k^* + 1) \geq 1$ , which means that the admissible value of  $\mu$  lies in the interval  $]c_{k^*}, c_{k^*+1}]$ . This further implies that the optimal number of strictly positive elements is  $k^*$ . We thus have to perform a linear search by computing  $L_1(k)$  and stopping when  $L_1(k^* + 1) \geq 1$ . If this last condition is never reached, that is,  $L_1(m) < 1$ , it means that all the  $m$  elements of  $\mathbf{p}$  are strictly positive and thus  $k^* = m$ .

### Computing the optimal probability distribution

Once the number of positive elements  $k^*$  is computed, the probability mass  $\mathbf{p}$  can be obtained from (41) in which  $s(k^*)$  is known after the evaluation of (43). The procedure is summarized as follows.

1. Sort and renumber the  $m$  elements by increasing cost ( $c_1 \leq c_2 \leq \dots \leq c_m$ ).

---

<sup>20</sup>Recall that the elements are sorted by increasing cost value.

2. Compute sequentially  $L_1(k)$  (Equation (43)) for  $k = 1, \dots, k^*$  until  $k^* = m$  or ( $L_1(k^*) < 1$  and  $L_1(k^* + 1) \geq 1$ ) (the  $\mu$  parameter is in the interval  $[k^*, k^* + 1]$ ).
3. Compute  $p_i$  by Equation (40) for  $i = 1, \dots, k^*$ .
4. Set  $p_i = 0$  for  $i = (k^* + 1), \dots, m$  if  $k^* < m$ .
5. Recover the initial numbering of the elements, that is, before executing step 1 (sorting).

## C.2 Results for the general $r > 1$ case and uniform reference probabilities

The case  $r \neq 2$  and  $r > 1$  is a bit more complex,

$$\begin{cases} \underset{\mathbf{p}}{\text{minimize}} & \mathbf{c}^T \mathbf{p} + \mathcal{Y} (\|\mathbf{p}\|_r)^r \\ \text{subject to} & \mathbf{e}^T \mathbf{p} = 1 \\ & \mathbf{p} \geq \mathbf{0} \end{cases} \quad (44)$$

where  $\|\cdot\|_r$  is the standard  $r$ -norm and we define  $\mathcal{Y} = T/(r-1)$  for convenience. This section proceeds similarly to the previous one and is based again on [85, 57]. As before, we assume that the costs  $c_i$  and the probabilities  $p_i$  are sorted by increasing value of cost.

### The Karush-Kuhn-Tucker conditions

By proceeding as in the previous Subsection C.1, the Karush-Kuhn-Tucker conditions are now

$$\begin{cases} \mathbf{c} + r\mathcal{Y}\mathbf{p}^{(r-1)} - \mu\mathbf{e} - \boldsymbol{\lambda} = \mathbf{0} \\ \boldsymbol{\lambda}^T \mathbf{p} = 0 \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \quad (45)$$

where  $(r)$  is the elementwise  $r$ -power. As before,  $\lambda_i p_i = 0$  for each  $i$ . A reasoning similar to the  $r = 2$  case provides the following extension of Equation (38)

$$p_i = \begin{cases} \sqrt[r-1]{\frac{1}{r\mathcal{Y}}(\mu - c_i)} & \text{when } \mu - c_i > 0 \\ 0 & \text{when } \mu - c_i \leq 0 \end{cases} \quad (46)$$

However, in this case, the Lagrange parameter  $\mu$  is difficult to find analytically so that the procedure derived in C.1 cannot be used. Kanzawa [57] therefore proposed to use a bisection method on the real line instead.

### A bisection procedure

As for the  $r = 2$  case, the main idea is to find numerically the optimal  $\mu$  by seeking the value which exactly satisfies the sum-to-one constraint  $\mathbf{e}^T \mathbf{p} = 1$ . Indeed, we observe from Equation (46) that each  $p_i$  taken independently is strictly increasing with respect to the  $\mu$  parameter when starting from the value  $\mu = c_i$  (thus in the interval  $[c_i, \infty[$ ) [57]. Moreover, from Equation (46),  $p_i \geq 0$ . This implies that the L1 norm of vector  $\mathbf{p}$  is strictly increasing from  $\mu = c_1$  (the minimum cost), and thus a bisection method (see, e.g., [93]) can be used in order to efficiently approximate this

quantity. The procedure is stopped when the L1 norm is sufficiently close to 1. Then, once  $\mu$  is closely approximated, Equation (46) is used in order to compute the probability mass  $\mathbf{p}$ .

Note that the author of [57] proposes, as initial lower and upper bounds for the admissible  $\mu$ ,

$$\begin{cases} \mu_{\text{inf}} &= c_1 \\ \mu_{\text{sup}} &= c_{\text{max}} + \frac{r\mathcal{Y}}{m^{r-1}} \end{cases} \quad (47)$$

where  $c_{\text{max}} = \max_{i \in \{1, \dots, m\}} \{c_i\} = c_m$  because  $\mathbf{c}$  is sorted and  $m$  is the last element of the vector. The lower bound is obvious. However, the upper bound might need a word of explanation. Let us show that this  $\mu_{\text{sup}}$  necessarily leads to a L1 norm of the corresponding  $\mathbf{p}$  vector greater or equal to 1. First, Equations (46) and (47) imply that each of the  $m$  elements of  $\mathbf{p}$  is strictly positive when using  $\mu_{\text{sup}}$ . Then, the second expression in (47) can be rearranged as

$$\frac{(\mu_{\text{sup}} - c_{\text{max}})}{r\mathcal{Y}} = \frac{1}{m^{r-1}}$$

which implies  $\sqrt[r-1]{\frac{1}{r\mathcal{Y}}(\mu_{\text{sup}} - c_{\text{max}})} = 1/m$ . Therefore, we must have  $\sqrt[r-1]{\frac{1}{r\mathcal{Y}}(\mu_{\text{sup}} - c_i)} \geq 1/m$  for each  $i$ . From (46), summing this expression over  $i = 1, \dots, m$  shows that the L1 norm of the corresponding  $\mathbf{p}$  vector is larger or equal to 1. The value  $\mu_{\text{sup}}$  in (47) is therefore an upper bound for the admissible  $\mu$  values. The bisection procedure follows

1. Sort and renumber the  $m$  elements by increasing cost ( $c_1 \leq c_2 \leq \dots \leq c_m$ ).
2. Initialize the lower bound and the upper bound of the  $\mu$  parameter as in Equation (47).
3. Perform a bisection search on  $\mu$  by computing  $\|\mathbf{p}\|_1$  from Equation (46) and testing if the result is lesser than or greater than 1 ( $\|\mathbf{p}\|_1$  is strictly increasing in function of  $\mu$ ). Stop when  $\|\mathbf{p}\|_1$  is sufficiently close to 1.
4. Compute  $\mathbf{p}$  from (46).
5. Recover the initial numbering of the elements, that is, before executing step 1 (sorting).

### C.3 Extension to arbitrary reference probabilities

We now start from the objective function defined in Equation (12), extending (45) by taking the reference probabilities into account,  $\mathbf{p}_{\text{ref}} = (p_i^{\text{ref}})$  (which correspond to  $\mathbf{p}_i^{\text{ref}}$  in (12)). The Karush-Kuhn-Tucker conditions become

$$\begin{cases} \mathbf{c} + r\mathcal{Y}(\mathbf{p} \div \mathbf{p}_{\text{ref}})^{(r-1)} - \mu\mathbf{e} - \boldsymbol{\lambda} = \mathbf{0} \\ \boldsymbol{\lambda}^T \mathbf{p} = 0 \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \quad (48)$$

where  $\mathcal{Y} = T/(r-1)$  and  $\div$  is the elementwise division. By a reasoning similar to the previous subsections (C.1-C.2), this leads to

$$p_i = \begin{cases} p_i^{\text{ref}} \sqrt[r-1]{\frac{1}{r\mathcal{Y}}(\mu - c_i)} & \text{when } \mu - c_i > 0 \\ 0 & \text{when } \mu - c_i \leq 0 \end{cases} \quad (49)$$

Notice that in this case, the elements are still sorted by increasing cost value, but then the  $p_i$  are no more ordered by decreasing value because they are modulated by  $p_i^{\text{ref}}$ .

As in Subsection C.2, a bisection procedure can be used in order to find the probability distribution  $\mathbf{p}$  summing to one. Following the derivation in Subsection C.2, the upper bound for the bisection procedure can be chosen here as  $\mu_{\text{sup}} = c_{\text{max}} + r\Upsilon$  because the reference probabilities  $p_i^{\text{ref}}$  sum to one. Thus the procedure for computing the probability distribution is exactly the same as in the previous subsection, except that Equation (49) is used instead of (46).

---

## References

- [1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network flows: theory, algorithms, and applications*. Prentice Hall, 1993.
- [2] T. Akamatsu. Cyclic flows, Markov process and stochastic traffic assignment. *Transportation Research B*, 30(5):369–386, 1996.
- [3] T. Akamatsu. Decomposition of path choice entropy in general transport networks. *Transportation Science*, 31(4):349–362, 1997.
- [4] 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 2011 conference*, pages 379–387. MIT Press, 2011.
- [5] K. Arrow, L. Hurwicz, and H. Uzawa. *Studies in linear and non-linear programming*. Stanford University Press, 1958.
- [6] A. L. Barabasi. *Network science*. Cambridge University Press, 2016.
- [7] F. Bavaud and G. Guex. Interpolating between random walks and shortest paths: A path functional approach. In K. Aberer, A. Flache, W. Jager, L. Liu, J. Tang, and C. Guéret, editors, *Proceedings of the 4th International Conference on Social Informatics (SocInfo '12)*, volume 7710 of *Lecture Notes in Computer Science*, pages 68–81. Springer, 2012.
- [8] A. Beck. *Introduction to nonlinear optimization*. SIAM, 2014.
- [9] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 2nd edition, 1999.
- [10] M. Blum, R. W. Floyd, V. R. Pratt, R. L. Rivest, and R. E. Tarjan. Time bounds for selection. *Journal of Computer and System Sciences*, 7(4):448–461, 1973.
- [11] I. Borg and P. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 1997.
- [12] U. Brandes and T. Erlebach, editors. *Network analysis: Methodological foundations*. Springer, 2005.
- [13] U. Brandes and D. Fleischer. Centrality measures based on current flow. In *Proceedings of the 22nd Annual Symposium on Theoretical Aspects of Computer Science (STACS '05)*, pages 533–544, 2005.



- [14] P. Buhlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer, 2011.
- [15] A. Busic and S. Meyn. Action-constrained Markov decision processes with kullback-leibler cost. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, pages 1431–1444. PMLR 75, 2018.
- [16] P. Chebotarev. A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. *Discrete Applied Mathematics*, 159(5):295–302, 2011.
- [17] P. Chebotarev. The walk distances in graphs. *Discrete Applied Mathematics*, 160(10–11):1484–1500, 2012.
- [18] 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.
- [19] 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.
- [20] P. Chebotarev and E. Shamis. On proximity measures for graph vertices. *Automation and Remote Control*, 59(10):1443–1459, 1998.
- [21] F. Chung and L. Lu. *Complex graphs and networks*. American Mathematical Society, 2006.
- [22] L. Condat. Fast projection onto the simplex and the  $\ell_1$  ball. *Mathematical Programming*, 158(1-2):575–585, 2016.
- [23] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, 3rd edition, 2009.
- [24] T. Cover and J. Thomas. *Elements of information theory*. Wiley, 2nd edition, 2006.
- [25] J. Culioli. *Introduction a l'optimisation*. Ellipses, 2012.
- [26] J.-C. Delvenne and A.-S. Libert. Centrality measures and thermodynamic formalism for complex networks. *Physical Review E*, 83(4):046117, 2011.
- [27] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.
- [28] A. Dolan and J. Aldous. *Networks and algorithms: An introductory approach*. Wiley, 1993.
- [29] P. G. Doyle and J. L. Snell. *Random walks and electric networks*. The Mathematical Association of America, 1984.
- [30] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the  $\ell_1$ -ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning (ICML '2008)*, pages 272–279, 2008.
- [31] E. Estrada. *The structure of complex networks*. Oxford University Press, 2012.
- [32] E. Estrada and N. Hatano. Communicability in complex networks. *Physical Review E*, 77(3):036111, 2008.

- [33] 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.
- [34] F. Fouss, M. Saerens, and M. Shimbo. *Algorithms and models for network data and link analysis*. Cambridge University Press, 2016.
- [35] R. Fox, A. Pakman, and N. Tishby. G-learning: taming the noise in reinforcement learning via soft updates. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2016)*, pages 202–211, 2001.
- [36] K. Francoise, I. Kivimaki, A. Mantrach, F. Rossi, and M. Saerens. A bag-of-paths framework for network data analysis. *Neural Networks*, 90:90–111, 2017.
- [37] A. L. Fred and A. K. Jain. Robust data clustering. In *Proceedings of the 2003 IEEE International Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’03)*, volume 2, pages 128–133, 2003.
- [38] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.
- [39] 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.
- [40] S. García-Díez, E. Vandembussche, and M. Saerens. A continuous-state version of discrete randomized shortest-paths. In *Proceedings of the 50th IEEE International Conference on Decision and Control (CDC ’11)*, pages 6570–6577, 2011.
- [41] M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized markov decision processes. In *Proceedings of the International Conference on Machine Learning (ICML 2019)*, pages 2160–2169, 2019.
- [42] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the USA*, 99(12):7821–7826, 2002.
- [43] I. Griva, S. Nash, and A. Sofer. *Linear and nonlinear optimization*. SIAM, 2nd edition, 2008.
- [44] 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.
- [45] G. Guex, S. Courtain, and M. Saerens. Covariance and correlation kernels on a graph in the generalized bag-of-paths formalism. *arXiv preprint arXiv:1902.03002 submitted for publication*, 2019.
- [46] G. Guex, I. Kivimaki, and M. Saerens. Randomized optimal transport on a graph: framework and new distance measures. *Network Science*, 7(1):88–122, 2019.
- [47] T. Hashimoto, Y. Sun, and T. Jaakkola. From random walks to distances on unweighted graphs. In *Advances in Neural Information Processing Systems 24: Proceedings of the NIPS ’15 Conference*, 2015.

- [48] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference, and prediction*. Springer, 2009.
- [49] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity*. CRC Press, 2015.
- [50] H. Havrda and F. Charvat. Quantification method of classification processes. concept of structural  $\alpha$ -entropy. *Kybernetika*, 3(1):30–35, 1967.
- [51] T. Hazan, R. Hardoon, and A. Shashua. Plsa for sparse arrays with Tsallis pseudo-additive divergence: noise robustness and algorithm. In *Proceedings of the 11th IEEE International Conference on Computer Vision*, pages 1–8. IEEE, 2007.
- [52] T. Hazan and A. Shashua. An efficient algorithm for maximum Tsallis entropy using fenchel-duality. Technical Report TR-110, The Hebrew University of Jerusalem, Israel, 2007.
- [53] M. Herbster and G. Lever. Predicting the labelling of a graph via minimum p-seminorm interpolation. In *Proceedings of the 22nd Conference on Learning Theory (COLT '09)*, pages 18–21, 2009.
- [54] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.
- [55] V. Ivashkin and P. Chebotarev. Do logarithmic proximity measures outperform plain ones in graph clustering? In *International Conference on Network Analysis*, pages 87–105. Springer, 2016.
- [56] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.
- [57] Y. Kanzawa. Generalization of quadratic regularized and standard fuzzy c-means clustering with respect to regularization of hard c-means. In V. Torra, Y. Narukawa, G. Navarro-Arribas, and D. Megías, editors, *Modeling Decisions for Artificial Intelligence*, pages 152–165, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- [58] Y. Kanzawa. Q-divergence-based relational fuzzy c-means clustering. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 22(1):34–43, 2018.
- [59] H. J. Kappen, V. Gómez, and M. Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.
- [60] J. N. Kapur. *Maximum-entropy models in science and engineering*. Wiley, 1989.
- [61] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.
- [62] C. Keylock. Simpson diversity and the shannon-wiener index as special cases of a generalized entropy. *Oikos*, 109(1):203–207, 2005.
- [63] 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.

- [64] D. J. Klein and M. Randić. Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95, 1993.
- [65] E. D. Kolaczyk. *Statistical analysis of network data: Methods and models*. Springer Series in Statistics. Springer, 2009.
- [66] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning (ICML '02)*, pages 315–322, 2002.
- [67] A. Laha, S. A. Chemmengath, P. Agrawal, M. Khapra, K. Sankaranarayanan, and H. Ramaswamy. On controllable sparse alternatives to softmax. In *Advances in Neural Information Processing Systems 32: Proceedings of the NeurIPS '18 Conference*, pages 6422–6432, 2018.
- [68] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.
- [69] K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Machine Learning Conference (ML95)*, pages 331–339, 1995.
- [70] K. Lee, S. Choi, and S. Oh. Maximum causal Tsallis entropy imitation learning. In *Advances in Neural Information Processing Systems 31: Proceedings of the NIPS 2010 Conference*, pages 4403–4413, 2018.
- [71] K. Lee, S. Choi, and S. Oh. Sparse Markov decision processes with causal sparse Tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018.
- [72] T. Lewis. *Network science*. Wiley, 2009.
- [73] Y. Li, Z.-L. Zhang, and D. Boley. The routing continuum from shortest-path to all-path: A unifying theory. In *Proceedings of the 31st International Conference on Distributed Computing Systems (ICDCS '11)*, pages 847–856. IEEE Computer Society, 2011.
- [74] Y. Li, Z.-L. Zhang, and D. Boley. From shortest-path to all-path: The routing continuum theory and its applications. *IEEE Transactions on Parallel and Distributed Systems*, 25(7):1745–1755, 2013.
- [75] D. G. Luenberger. *Introduction to dynamic systems: Theory, models, and applications*. Wiley, 1979.
- [76] D. G. Luenberger and Y. Ye. *Linear and nonlinear programming*. Springer, 3rd edition, 2010.
- [77] D. Lusseau. The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(suppl\_2):S186–S188, 2003.
- [78] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.
- [79] 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.

- [80] C. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.
- [81] A. Martins and R. Astudillo. From softmax to sparsemax: a sparse model of attention and multi-label classification. In *Proceedings of the International Conference on Machine Learning (ICML-2016)*, pages 1614–1623, 2016.
- [82] M. Menard, V. Courboulay, and P.-A. Dardignac. Possibilistic and probabilistic fuzzy clustering: unification within the framework of the non-extensive thermostatics. *Pattern Recognition*, 36(6):1325–1342, 2003.
- [83] M. Minoux. *Mathematical programming, theory and algorithms*. John Wiley, 1986.
- [84] S. Miyamoto, H. Ichihashi, and K. Honda. *Algorithms for fuzzy clustering*. Springer, 2008.
- [85] S. Miyamoto and K. Umayahara. Fuzzy clustering by quadratic regularization. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pages 1394–1399, 1998.
- [86] B. Muzellec, R. Nock, G. Patrini, and F. Nielsen. Tsallis regularized optimal transport and ecological inference. In *Proceedings of the 31 International Conference of the Association for the Advancement of Artificial Intelligence (AAAI 2017)*, 2017.
- [87] M. E. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.
- [88] M. E. J. Newman. *Networks: An introduction, 2nd ed.* Oxford University Press, 2018.
- [89] C. Ngyen and H. Mamitsuka. New resistance distances with global information on large graphs. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS '16)*, pages 639–647, 2016.
- [90] J. R. Norris. *Markov chains*. Cambridge University Press, 1997.
- [91] M. Panzacchi, B. Van Moorter, O. Strand, M. Saerens, I. Kivimaki, C. St Clair, I. Herfindal, and L. Boitani. Predicting the continuum between corridors and barriers to animal movements using step selection functions and randomized shortest paths. *Journal of Animal Ecology*, 85(1):32–42, 2016.
- [92] L. Peliti. *Statistical mechanics in a nutshell*. Princeton University Press, 2011.
- [93] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical recipes: The art of scientific computing*. Cambridge University Press, 3rd edition, 2007.
- [94] W. L. Price. *Graphs and networks: an introduction*. London Butterworths, 1971.
- [95] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
- [96] R. Rardin. *Optimization in operations research*. Prentice Hall, 1998.
- [97] L. E. Reichl. *A modern course in statistical physics*. Wiley, 2nd edition, 1998.

- [98] J. Rubin, O. Shamir, and N. Tishby. *Trading value and information in MDPs*, pages 57–74. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [99] M. Saerens, Y. Achbany, F. Fouss, and L. Yen. Randomized shortest-path problems: Two related models. *Neural Computation*, 21(8):2363–2404, 2009.
- [100] B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press, 2002.
- [101] T. Silva and L. Zhao. *Machine learning in complex networks*. Springer, 2016.
- [102] F. Sommer, F. Fouss, and M. Saerens. Comparison of graph node distances on clustering tasks. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN 2016)*. *Lecture Notes in Computer Science*, volume 9886, pages 192–201, 2016. Springer.
- [103] F. Sommer, F. Fouss, and M. Saerens. Modularity-driven kernel k-means for community detection. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN 2017)*. *Lecture Notes in Computer Science*, volume 10614, pages 423–433, 2017. Springer.
- [104] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.
- [105] L. Tang and H. Liu. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*, pages 817–826, 2009.
- [106] 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 '09)*, pages 1107–1116, 2009.
- [107] L. Tang and H. Liu. Toward predicting collective behavior via social dimension extraction. *IEEE Intelligent Systems*, 25(4):19–25, 2010.
- [108] H. M. Taylor and S. Karlin. *An introduction to stochastic modeling*. Academic Press, 3rd edition, 1998.
- [109] P. D. Taylor. Inclusive fitness arguments in genetic models of behaviour. *Journal of Mathematical Biology*, 34(5–6):654–674, 1996.
- [110] M. Thelwall. *Link analysis: An information science approach*. Elsevier, 2004.
- [111] E. A. Theodorou, D. Krishnamurthy, and E. Todorov. From information theoretic dualities to path integral and kullback-leibler control: continuous and discrete time formulations. In *The Sixteenth Yale Workshop on Adaptive and Learning Systems*, 2013.
- [112] E. A. Theodorou and E. Todorov. Relative entropy and free energy dualities: Connections to path integral and kl control. In *Proceedings of the 51st IEEE Conference on Decision and Control (CDC 2012)*, pages 1466–1473. IEEE, 2012.
- [113] E. Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, pages 1369–1375. MIT Press, 2007.

- [114] E. Todorov. General duality between optimal control and estimation. In *Proceedings of 47th IEEE Conference on Decision and Control (CDC'08)*, pages 4286–4292, 2008.
- [115] C. Tsallis. Generalized entropy-based criterion for consistent testing. *Physical Review E*, 58(2):1442, 1998.
- [116] C. Tsallis. *Introduction to nonextensive statistical mechanics*. Springer, 2009.
- [117] 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, 2010.
- [118] U. von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15:1751–1798, 2014.
- [119] W. Wang and M. Carreira-Perpinan. Projection onto the probability simplex: an efficient algorithm with a simple proof, and an application. *ArXiv preprint arXiv:1309.1541 [cs.LG]*, 2013.
- [120] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 1994.
- [121] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.
- [122] 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 '07)*, volume 4426 of *Lecture Notes in Artificial Intelligence*, pages 1037–1045. Springer, 2007.
- [123] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens. Graph nodes clustering with the sigmoid commute-time kernel: A comparative study. *Data & Knowledge Engineering*, 68(3):338–361, 2009.
- [124] 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 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*, pages 785–793, 2008.
- [125] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.
- [126] 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.
- [127] 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.