Data Mining and Knowledge Discovery (2021) 35:986–1031
https://doi.org/10.1007/s10618-021-00742-y

Sparse randomized shortest paths routing with Tsallis divergence regularization

Pierre Leleux[1] · Sylvain Courtain[1] · Guillaume Guex[2] · Marco Saerens[3,4]

Received: 7 May 2020 / Accepted: 28 January 2021 / Published online: 6 March 2021

Abstract

This work elaborates on the important problem of (1) designing optimal randomized routing policies for reaching a target node t from a source note 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 positive temperature parameter T. To this end, the randomized shortest path (RSP) formalism 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.

Responsible editor: Hanghang Tong.

✉ Pierre Leleux
p.leleux@uclouvain.be

Sylvain Courtain
sylvain.courtain@uclouvain.be

Guillaume Guex
guillaume.guex@unil.ch

Marco Saerens
marco.saerens@uclouvain.be

1 LouRIM, Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium

2 SLI, Université de Lausanne, Lausanne, Switzerland

3 ICTEAM, Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium

4 IRIDIA, Université Libre de Bruxelles, Bruxelles, Belgium

Keywords Graph mining · Link analysis · Network data analysis · Network science · Distances between nodes

1 Introduction

1.1 General introduction

Graph mining and link analysis are currently used in a large number of different fields for analyzing network data, like social networks, networks of transactions, protein networks, road networks, etc. (e.g. Barabasi 2016; Brandes and Erlebach 2005; Chung and Lu 2006; Estrada 2012; Fouss et al. 2016; Kolaczyk 2009; Lewis 2009; Newman 2018; Silva and Zhao 2016; Thelwall 2004; Wasserman and Faust 1994). 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 Sect. 1.2), first developed in the transportation sciences (Akamatsu 1996) and then further extended in the context of network data analysis (Françoisse et al. 2017; Kivimäki et al. 2014; Saerens et al. 2009; Yen et al. 2008). The intuition behind this model is as follows. 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[1] from s to t, depending on a temperature parameter T. The agent then chooses a path 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 Akamatsu (1997), Bavaud and Guex (2012), Saerens et al. (2009), 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 the KL divergence to be replaced by the Tsallis divergence (Tsallis 1998, 2009), also called the Havrda-Charvat divergence (Havrda and Charvat 1967; Kapur 1989), with the consequence that the obtained policy becomes *sparse* for low temperatures. 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,

[1] Considered as independent.

the most costly paths are gradually removed until only the shortest paths remain active (see the illustration in Fig. 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 Kanzawa (2013), Miyamoto and Umayahara (1998), and Saerens et al. (2009) 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.[2] 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 (Kivimäki et al. 2014). Both capture the notion of *relative accessibility* (proximity and amount of inter-connectivity (Chebotarev and Shamis 1998)) 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 (Guex et al. 2019; Lebichot et al. 2018): 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 (Freeman 1977) are defined from least-cost paths, whereas resistance distance and random walk centrality (Brandes and Fleischer 2005; Newman 2005) rely on random walks (Doyle and Snell 1984). However, in practice, movements over a network hardly ever occur either perfectly optimally or perfectly randomly: the agent usually has some, albeit incomplete, knowledge of his environment. As already mentioned, the RSP framework (Françoisse et al. 2017; Kivimäki et al. 2014; Saerens et al. 2009; Yen et al. 2008) 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., to model the behavior of animals during migration (Panzacchi et al. 2016). Initially, the model was inspired by transportation models developed in the transportation sciences (Akamatsu 1996), and is also closely related to the work of Todorov (2007, 2008) in reinforcement learning and Delvenne and Libert (2011) in defining random walks on a graph with maximal entropy or minimal free energy.

Besides this previous work, 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 (Alamgir and von Luxburg 2011; Herbster and Lever 2009; Ngyen and Mamitsuka 2016; von Luxburg et al. 2014), others on combinatorial analysis arguments (Chebotarev 2011,

[2] We conjecture that the Tsallis FE is a distance measure, as it verified the properties of a distance measure in all our applications.

2012, 2013), on L1–L2 regularization (Li et al. 2011, 2013), as well as on network flow models with Kullback–Leibler divergence regularization (Bavaud and Guex 2012; Guex and Bavaud 2015; Guex 2016). These last propositions are closely related 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 (von Luxburg et al. 2010, 2014) 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 (Fouss et al. 2007)]. For a more thorough discussion of this related work, see, e.g., Fouss et al. (2016) or Françoisse et al. (2017).

Among these studies, the most closely related work is from Li et al. (2011, 2013). The authors present a sparse routing strategy based on edge flow optimization regularized by mixed L1/L2 norms (see for instance Hastie et al. 2015). More precisely, they use the edge flow formalism (Ahuja et al. 1993; Dolan and Aldous 1993) 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 great interest, there are some significant differences. First, they consider undirected networks, while we address directed ones. They also consider net flows (as in electrical networks), while the present work considers raw flows. Moreover, the regularization technique is different: in Li et al. (2011, 2013), a mixed L1/L2 norm is minimized, whereas we use a Tsallis divergence regularization term.[3] 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 work of Li et al. (2011, 2013) is based on standard network flow theory. Furthermore, one of our main goals is to derive dissimilarity measures between nodes, which is not addressed in Li et al. (2011, 2013).

Concerning Tsallis entropy regularization, it has been known for years that when minimizing the regularized expected cost together with sum-to-one and non-negativity constraints, the resulting solution becomes sparse when increasing the regularization parameter. 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 (Buhlmann and van de Geer 2011; Hastie et al. 2009, 2015). For instance, it was shown by Kanzawa (2013) and Miyamoto and Umayahara (1998) 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 by Menard et al. (2003) and Miyamoto et al. (2008); however, Menard et al. (2003) consider a different objective function which does not lead to sparse solutions. Another interesting application is the sparse probabilistic latent semantic analysis introduced by Hazan et al. (2007). The authors propose an expectation-maximization algorithm based on Tsallis divergence, instead of KL divergence. In another paper (Hazan and Shashua 2007), some of 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 (Hazan and Shashua 2007, Subsection 1.2), with only one of these ways

[3] See the discussion following Eq. (10) for details about the differences between the two objective functions.

being relevant in our applications. Note also that Kanzawa (2018) 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 of Lee et al. (2018b) investigates sparse policies for discounted **Markov decision processes** (MDP), with application to reinforcement learning (see also Geist et al. 2019 generalizing these initiatives into a more general framework). As initiated by Todorov (2007, 2008) (and related to the alternative view of the RSP presented in Sect. 2.3), the authors add an entropy term to the cost associated with each action choice in the MDP. Instead of using the KL divergence as in Todorov (2007), they choose the Tsallis entropy with parameter[4] $r = 2$ and a uniform reference probability distribution [see Eq. (8) below]. As in Kanzawa (2013) and Miyamoto and Umayahara (1998), they show that the resulting policy enforces sparsity. This problem is in fact closely connected to the recently introduced sparse maximum procedure (Laha et al. 2018; Martins and Astudillo 2016) and the probability simplex projection, e.g. Beck (2014), Condat (2016), Duchi et al. (2008), and Wang and Carreira-Perpinan (2013), who proposed procedures related to Kanzawa (2013) and Miyamoto and Umayahara (1998) for solving this problem. Our work therefore solves a problem similar to Lee et al. (2018b), 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 Eq. (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 future work. Note also that we are working with absorbing Markov chains, while Lee et al. (2018b) investigate 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 and study 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 sparse outputs of a classifier by Martins and Astudillo (2016) in the context of multi-label classification, and Lee et al. (2018a, b) proposing a "maximum causal Tsallis entropy" framework (see Lee et al. (2018a) for details). Finally Muzellec et al. (2017) 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 analyze the optimization problem by using the r-exponential formalism (Tsallis 2009). 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 as follows:

[4] We will see later that the Tsallis entropy and divergence both depend on an r parameter influencing the shape of the resulting probability mass; see Eq. (8) and the discussion after Eq. (12).

- 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. This extends previous results by minimizing the Tsallis divergence instead of the Tsallis entropy.
- Two new dissimilarity measures between nodes, interpolating[5] between the least-cost and the commute cost distances based on a model of sparse movements on the network, are introduced.
- 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 is provided.

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 Sect. 3, the sparse RSP model of movement is developed based on Tsallis divergence regularization. Section 4 introduces the two new dissimilarity measures between nodes. Experiments on node clustering and semi-supervised classification tasks are presented in Sect. 5. Finally, Sect. 6 concludes the work.

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 (Sect. 2.3) and then developing its sparse version (Sect. 3).

2.1 Background and notation

Let us first introduce some background and notation. Consider a weighted, directed,[6] 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 with each node are set proportionally to the affinities of the incident edges and then normalized in order to sum to one. Alternatively, they can be chosen as uniform, depending on the problem,

[5] For an undirected graph and up to a scaling factor.

[6] 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).

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

where $Succ(i)$ is the set of successor nodes of node i and $|Succ(i)|$ 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 with 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 large value (denoted by ∞) and the products $a_{ij}c_{ij} = p_{ij}c_{ij} = 0$ by convention. Usually, costs are set independently of the affinities, depending on the application, but, if there are no obvious cost values related to the problem, we can, e.g., set $c_{ij} = 1$ or $c_{ij} = 1/a_{ij}$ as in electric networks (see Françoisse et al. 2017 for a discussion).

A path $\wp$ 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 goal node is transformed into an *absorbing* and *killing* node. That is, when the node is reached, the random walker 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 (Fouss et al. 2016) where the target node has no outgoing edge. Thus, the target node has no successor node, $\mathcal{S}ucc(t) = \varnothing$, on this new graph.

The *total cost* of a path, $\tilde{c}(\wp)$, is simply the sum of the edge costs c_{ij} along $\wp$, while the *length* of a path $\ell(\wp)$ (or simply ℓ) 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 (Saerens et al. 2009), which comes from the transportation sciences (Akamatsu 1996), was further developed by exploiting basic concepts from statistical physics (Françoisse et al. 2017; Kivimäki et al. 2014; Saerens et al. 2009; Yen et al. 2008). A similar model was proposed in reinforcement learning and process control by Todorov (2007, 2008), from a different 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[7]) instead of standard "local" flows defined on nodes and edges (the "micro" level) (Ahuja et al. 1993; Dolan and Aldous 1993). Among other properties, it provides an optimal, randomized routing policy from source node s to target node t based on the minimization over the set of paths $\wp \in \mathcal{P}_{st}$ of the (relative) KL-based **free energy** (FE) of statistical physics (Jaynes 1957; Peliti 2011; Reichl 1998),

[7] For a generalization to several input-outputs, see the work of Guex et al. (2019).

$$\begin{aligned} &\underset{\{\mathrm{P}(\wp)\}_{\wp\in\mathcal{P}_{st}}}{\text{minimize}}\ \phi_{st}^{\mathrm{KL}}(\mathrm{P}) = \underbrace{\sum_{\wp\in\mathcal{P}_{st}} \mathrm{P}(\wp)\,\tilde{c}(\wp)}_{\text{expected cost}} + T \underbrace{\sum_{\wp\in\mathcal{P}_{st}} \mathrm{P}(\wp)\log\left(\frac{\mathrm{P}(\wp)}{\tilde{\pi}(\wp)}\right)}_{\text{KL regularization}} \\ &\text{subject to } \sum_{\wp\in\mathcal{P}_{st}} \mathrm{P}(\wp) = 1 \end{aligned} \tag{2}$$

where $\tilde{c}(\wp) = \sum_{\tau=1}^{\ell} c_{s(\tau-1)s(\tau)}$ is the total cumulated cost along path $\wp$ when visiting the sequence of nodes, or states, $(s(\tau))_{\tau=0}^{\ell}$ in the sequential order, and ℓ is the length of path $\wp$. Furthermore, $\tilde{\pi}(\wp) = \prod_{\tau=1}^{\ell} p^{\mathrm{ref}}_{s(\tau-1)s(\tau)}$ is the random walk probability of the path, that is, the product of the reference transition probabilities (1) along hitting path $\wp$ ending in (killing and absorbing) target node t. The objective function (2) 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 the target node from the 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 Eq. (1)].

As is well known, this free energy minimization problem, akin to maximum entropy models (Cover and Thomas 2006; Jaynes 1957; Kapur 1989), leads to a *Gibbs–Boltzmann* probability distribution on the set of paths (see, e.g., Françoisse et al. 2017),

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

where $\theta = 1/T$ is the inverse temperature and the denominator $\mathcal{Z} = \sum_{\wp\in\mathcal{P}_{st}} \tilde{\pi}(\wp)\exp[-\theta\tilde{c}(\wp)]$ 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[8]

$$p^*_{ij} = p^{\mathrm{ref}}_{ij}\exp\left[-\theta(c_{ij} + \phi^{\mathrm{KL}}_{jt}(\mathrm{P}^*) - \phi^{\mathrm{KL}}_{it}(\mathrm{P}^*))\right] \text{ for all } i \neq t \tag{4}$$

and $p^*_{tj} = 0$ for target node t and all j. The free energy between i and t at the optimal probability distribution,[9] $\phi^{\mathrm{KL}}_{it}(\mathrm{P}^*)$, can be computed by solving a system of linear equations of size n (Françoisse et al. 2017; Kivimäki et al. 2014; Saerens et al. 2009). As already mentioned, this defines a *biased* random walk on G where the

[8] Here, we used $\phi^{\mathrm{KL}}_{it}(\mathrm{P}^*) = -\frac{1}{\theta}\log[z_{it}]$ and $p^*_{ij} = p^{\mathrm{ref}}_{ij}\exp[-\theta c_{ij}]z_{jt}/z_{it}$ with $z_{ij} = [\mathbf{Z}]_{ij} = [(\mathbf{I} - \mathbf{W})^{-1}]_{ij}$ and $[\mathbf{W}]_{ij} = w_{ij} = p^{\mathrm{ref}}_{ij}\exp[-\theta c_{ij}]$ (see Françoisse et al. (2017), Kivimäki et al. (2014) and Saerens et al. (2009) for details).

[9] Providing the directed free energy distance (Kivimäki et al. 2014).

random walker is more and more "attracted" by the target node t when T decreases. Interestingly, the policy is independent of the source node s. This also implies that for any path $\wp$ visiting nodes $s(\tau)$, $\tau = 0, \ldots, \ell(\wp)$, we have $\mathrm{P}^*(\wp) = \prod_{\tau=1}^{\ell} p^*_{s(\tau-1)s(\tau)}$.

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 Akamatsu 1997; Bavaud and Guex 2012; Saerens et al. 2009) 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 the standard procedures developed in Françoisse et al. (2017), Kivimäki et al. (2014) and Saerens et al. (2009), which are quite efficient. However, the advantage is that this algorithm can be easily adapted to other divergence regularizations.

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

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

where $\mathcal{P}red(j)$ is the set of predecessor nodes of node j and δ_{sj} is a Kronecker delta corresponding to the unit input flow in source node s. This formulation of the optimization problem is derived from Eq. (A.2), Appendix A.1, and then solved in Appendix A.2. In this Eq. (5), the p_{ij} are the elements of the transition matrix $\mathbf{P}$ to be found (the *routing policy*), the p^{ref}_{ij} are the elements of the transition matrix of the reference random walk on the graph [see Eq. (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., Norris 1997; Taylor and Karlin 1998). 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), we have $\bar{n}_t = 1$. Therefore, row t of the transition matrix[10] $\mathbf{P}$ contains only zeros and the matrix is sub-stochastic.

Following the Lagrange formulation of the problem appearing in Eq. (A.3) and inspired by discrete space-time optimal control (see, e.g., Bryson and Ho 1975; Luenberger 1979), the p_{ij} can be regarded as decision/control variables (usually denoted as $\mathbf{u}$ in control theory) and the $\bar{n}_j$ as state values (denoted as $\mathbf{x}$ in control theory), and

[10] Defining a killed random walk on the graph because node t is absorbing and killing.

are considered as independent [the relations between these variables are encoded in the (linear) constraints]. The state values are uniquely defined for a given value of the control variables. 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 Eq. (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 Appendix A [see Eqs. (A.4) and (A.6)], it is shown that the transition probabilities, providing the routing policy, can be computed by sequentially updating the Lagrange parameters λ_i^{KL} (associated with the computation of the expected number of visits to nodes) from the transition probabilities and vice versa.

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

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

– Compute the transition probabilities:

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

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

Note that because the target node t is absorbing and killing, Eq. (6) provides $\lambda_t^{\mathrm{KL}} = 0$. We also observe that, for missing edges, the transition probabilities are equal to zero, as they should be. Interestingly, after convergence, the Lagrange parameters are nothing other than the *optimal directed free energy distance*, $\lambda_i^{\mathrm{KL}} = \phi_{it}^{\mathrm{KL}}(\mathrm{P}^*)$, appearing in Eq. (2) and evaluated at the Gibbs–Boltzmann distribution P^* [Eq. (3)], as discussed in Appendix A.2. Moreover, notice that, for an undirected graph and edge costs defined as $c_{ij} = 1/a_{ij}$, Eq. (6) is nothing more than the harmonic function (induced by Kirchhoff's current law and Ohm's law) computing the voltage associated with the nodes, when considering sources of voltage on nodes (Doyle and Snell 1984).

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. An illustrative example of this property is provided at the end of this section (see Fig. 2).

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 $p_j^{\text{ref}} > 0$, the Tsallis directed r-divergence (Tsallis 1998, 2009; see also Havrda and Charvat 1967; Kapur 1989) 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 Eq. (12). This measure (8) generalizes the KL divergence (Kapur 1989) in the sense that it converges to this quantity when $r \to 1^+$. 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., Kapur 1989; Keylock 2005). 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 Eq. (5), the aim is to minimize the Tsallis-based free energy function instead of the KL-based free energy of Eq. (2),

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

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.[11] The Lagrange function [see Appendix A.2,

[11] It is assumed that the source node is different from the target node and that the target node n has no successor node.

especially Eq. (A.3), for details] integrating the equality constraints (but not the non-negativity constraints restricting the domain of $\mathbf{P}$ and $\bar{\mathbf{n}}$) becomes

$$\mathscr{L}(\mathbf{P}, \bar{\mathbf{n}}; \boldsymbol{\mu}, \boldsymbol{\lambda}_{\mathrm{Ts}}) = \sum_{i \in \mathcal{V} \backslash n} \bar{n}_i \sum_{j \in \mathcal{S}ucc(i)} p_{ij} \left(c_{ij} + \tfrac{T}{r-1} \left(\left(\frac{p_{ij}}{p_{ij}^{\mathrm{ref}}} \right)^{r-1} - 1 \right) \right) + \sum_{j \in \mathcal{V}} \lambda_j^{\mathrm{Ts}} \left(\sum_{i \in \mathcal{P}red(j)} \bar{n}_i p_{ij} + \delta_{1j} - \bar{n}_j \right) + \sum_{i \in \mathcal{V} \backslash n} \mu_i \left(1 - \sum_{j \in \mathcal{S}ucc(i)} p_{ij} \right) \tag{10}$$

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, $\bar{n}_i \in \mathbb{R}_+$. There are n Lagrange parameters λ_j^{Ts} associated with the nodes and gathered in column vector $\boldsymbol{\lambda}_{\mathrm{Ts}}$. As for the standard RSP, for $\theta = 1/T \to 0^+$, this model becomes a simple absorbing random walk on G with transition probabilities provided by Eq. (1) (reference probabilities). For $\theta \to \infty$, it provides least-cost routing concentrated on shortest paths.

Let us proceed as in the 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.

Before we proceed, a remark concerning related work. If we compare the objective function used in (10) with the one studied by Li et al. (2011, 2013) [see Li et al. 2013, Eq. (18)—which also leads 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 different. Indeed, the objective function of Li et al. (2011, 2013) 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 also that Li et al. (2011, 2013) consider a symmetric cost matrix and net flows, instead of raw flows as in the present paper.

As for the previous section, and as detailed in Appendix A, our optimization procedure optimizes sequentially the objective function by Lagrange duality (Culioli 2012; Griva et al. 2008; Minoux 1986). 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 λ_i^{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 (Bertsekas 1999). A discussion of the convexity of the objective function with respect to the edge flows appears in Appendix B. In short, it is shown by heuristic arguments that the objective function appearing in Eq. (9) is convex with respect to the edge flows $\bar{n}_{ij}$, although a formal proof is left for future work—it therefore remains a conjecture at this stage. Then, by using the same reasoning as for the KL divergence case of Appendix 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 from the previous section (KL regularization), namely that the non-negativity of the transition probabilities is no longer

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 rearrange the Lagrange function (10) after keeping all the terms depending on the transition probabilities and using $\sum_{(i,j)\in\mathcal{E}} p_{ij}x_{ij} = \sum_{i\in\mathcal{V}\setminus n}\sum_{j\in\mathcal{S}ucc(i)} p_{ij}x_{ij} = \sum_{i=1}^{n-1}\sum_{j=1}^{n} p_{ij}x_{ij} = \sum_{j\in\mathcal{V}}\sum_{i\in\mathcal{P}red(j)} p_{ij}x_{ij}$ for any quantity x_{ij} defined on the edges of our network (recall that $p_{ij} = 0$ for missing edges and that node 1 is the source node and node n the target node),

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

because $\bar{n}_n = 1, \bar{n}_\bullet = \sum_{i=1}^{n} \bar{n}_i$ and $\sum_{j=1}^{n} p_{ij} = 1$ when $i \neq n$. In this last expression, we defined the *augmented costs* for each row i [and thus each transient (non-absorbing) node i], $c'_{ij} = c_{ij} + \lambda_j^{\mathrm{Ts}}$.

From (11), we observe that the part of the objective function depending on the transition probabilities (first term of first line) is a sum of $(n-1)$ non-negative 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 $i = 1, \ldots, (n-1)$ in turn is

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

together with

$$\begin{cases} \mathbf{p}_i^{\mathrm{ref}} = \mathbf{row}_i(\mathbf{P}_{\mathrm{ref}}) \\ \mathbf{c}'_i = \mathbf{row}_i(\mathbf{C} + \mathbf{e}\boldsymbol{\lambda}_{\mathrm{Ts}}^{\mathrm{T}}) \\ r > 1 \end{cases}$$

where $\mathbf{row}_i(\mathbf{P}_{\text{ref}})$ extracts the values of the successor nodes of node i (for which $p_{ij}^{\text{ref}} > 0$) from the ith row of matrix $\mathbf{P}_{\text{ref}}$. Therefore, only the values corresponding to existing links are extracted—missing links are ignored, as should be. In this last expressions, $\div$ 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. Note that the augmented costs are only defined on existing edges (with strictly positive reference transition probabilities). Elementwise, this provides

$$c'_{ij} = c_{ij} + \lambda_j^{\text{Ts}} \quad \text{with } (i,j) \in \mathcal{E}$$

This shows that, as in many other cases, the Lagrange parameters λ_j^{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 Eq. (12).

Note that the r parameter controls the impact of the costs on the probabilities. For example, for uniform reference probabilities and linear costs, $\mathbf{c} = [1,2,3,4,5]^{\text{T}}$, the form of the resulting probability distribution will be decreasing and convex when $r < 2$ ($\mathbf{p} = [0.480, 0.295, 0.156, 0.060, 0.009]^{\text{T}}$ when $r = 1.5$ and $T = 1$), linear when $r = 2$ ($\mathbf{p} = [0.4, 0.3, 0.2, 0.1, 0.0]^{\text{T}}$ with $T = 1$), and concave when $r > 2$ ($\mathbf{p} = [0.289, 0.262, 0.229, 0.182, 0.037]^{\text{T}}$ when $r = 4$ and $T = 1$). On the other hand, the temperature T mainly controls the sparseness of the distribution ($\mathbf{p} = [0.533, 0.333, 0.133, 0.0, 0.0]^{\text{T}}$ when $r = 2$ with $T = 0.5$ and $\mathbf{p} = [0.240, 0.220, 0.200, 0.180, 0.160]^{\text{T}}$ when $r = 2$ with $T = 5$).

Thus, the problem of computing the transition probabilities reduces to the minimization of a linear function with an 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 (Kanzawa 2013; Miyamoto and Umayahara 1998). 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 also that Miyamoto and Umayahara (1998) solve the quadratic regularization problem with $r = 2$, while Kanzawa (2013) 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 Appendix C for each node $i \neq t$,

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

which returns a possibly sparse discrete probability mass $\mathbf{p}_i^*$ for node i. In short, the solution for the general case is of the form [Eq. (C.13)]

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

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 1s, and the threshold μ_i must be chosen in order to satisfy the sum-to-one constraint, $(\mathbf{p}_i^*)^{\mathrm{T}}\mathbf{e} = 1$. Expression (14) is the counterpart of Eq. (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 $\boldsymbol{\lambda}_{\mathrm{Ts}}$ through the augmented costs.

Based on previous work from Kanzawa (2013) and Miyamoto and Umayahara (1998), three algorithms are detailed in Appendix 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 a more 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 (at least on sparse graphs). Our contribution with respect to Kanzawa (2013) and Miyamoto and Umayahara (1998) is the introduction of non-uniform reference probabilities in C.3 which allow us 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 $\boldsymbol{\lambda}_{\mathrm{Ts}}$, follows the same principle as for the KL divergence [see Eq. (6) and Appendix A.2 for details], the only difference being the definition of the divergence. Indeed, the elements of $\tilde{\mathbf{h}}_{\mathrm{KL}}$, based on the KL divergence, in Eq. (A.5), now become $\tilde{h}_i^{\mathrm{Ts}} = \frac{1}{r-1}\sum_{j\in\mathcal{S}ucc(i)} p_{ij}\left((p_{ij}/p_{ij}^{\mathrm{ref}})^{r-1} - 1\right)$ for the Tsallis divergence.

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

$$\boldsymbol{\phi}_{\mathrm{Ts}} \triangleq \boldsymbol{\lambda}_{\mathrm{Ts}} \tag{15}$$

and we verify that they correspond to the minimized free energy objective function (9). These quantities are obtained by solving the following system of linear equations. involving the sub-stochastic matrix $\mathbf{P}$,

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

with respect to $\boldsymbol{\lambda}_{\mathrm{Ts}}$. Note that these equations imply $\lambda_n^{\mathrm{Ts}} = 0$ for absorbing and killing target node n. In matrix form, we have

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

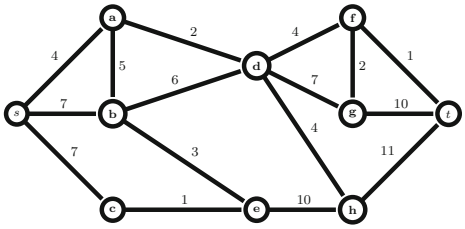

Fig. 1 A small undirected graph composed of 10 nodes with costs on edges; adapted from Price (1971)

where $\tilde{\mathbf{c}} = (\mathbf{P} \circ \mathbf{C})\mathbf{e}$ and $\tilde{\mathbf{h}}_{\text{Ts}} = \frac{1}{r-1}\left[\mathbf{P} \circ \left((\mathbf{P} \div \mathbf{P}_{\text{ref}})^{(r-1)} - \mathbf{E}\right)\right]\mathbf{e}$, and $\mathbf{e}$ 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. It can be observed that the Tsallis free energy can be interpreted as a kind of weighted averaged value of the objective function (9) before being absorbed by the target node,

$$\boldsymbol{\phi}_{\text{Ts}} = \boldsymbol{\lambda}_{\text{Ts}} = (\mathbf{I} + \mathbf{P})^{-1}(\tilde{\mathbf{c}} + T\tilde{\mathbf{h}}_{\text{Ts}}) = (\mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \cdots)(\tilde{\mathbf{c}} + T\tilde{\mathbf{h}}_{\text{Ts}})$$

Finally, after initializing the transition probabilities to the reference probabilities, the overall procedure for computing the optimal randomized policy based on Tsallis divergence aims to iterate Eqs. (17) and (13) until convergence. Each step has a unique optimal solution so that the iterative procedure converges (Bertsekas 1999). As before, we observed that, in all our runs, the duality gap is zero, showing that a global minimum is reached. After convergence, the algorithm provides an optimal, possibly sparse, routing policy taking the form of a Markov chain with absorbing, killing node n (or t before renumbering the nodes).

3.4 Illustrative example

In this subsection, we 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).

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

We illustrate how the parameter θ 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 Fig. 2.

As the value of θ 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 θ increases. This property is also observed for other values of the parameter r.

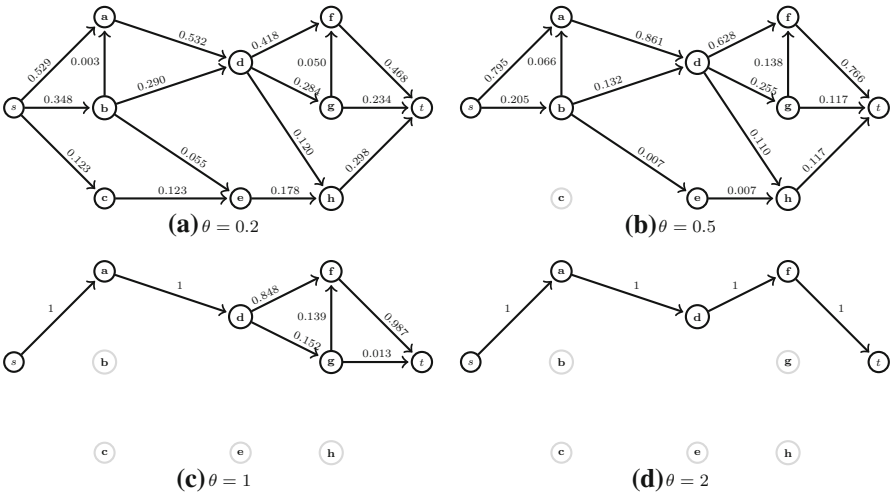

Fig. 2 Representation of the net flow from s to t in terms of θ for the Tsallis RSP and using $r = 2$

4 Derived dissimilarities

We now define two new dissimilarity measures between nodes interpolating between the least-cost and the commute cost,[12] 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 (Françoisse et al. 2017; Kivimäki et al. 2014; Yen et al. 2008), the quantity ϕ_{it}^{Ts}, provided by the Lagrange parameter λ_i^{Ts} after convergence [see Eq. (15)], is called the Tsallis directed FE from any node i to absorbing, killing node[13] t. Accordingly, the *Tsallis FE distance* is the symmetrized quantity

$$\Delta_{st}^{\mathrm{TsFE}} = \frac{\phi_{st}^{\mathrm{Ts}} + \phi_{ts}^{\mathrm{Ts}}}{2} \tag{18}$$

Interestingly, we found that this quantity satisfies the triangle inequality on all the investigated datasets and values of the θ 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 (Françoisse et al. 2017)) and hope to prove this 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

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

[12] And thus also the resistance distance based on the effective resistance when c_{ij} is defined as $1/a_{ij}$, as in electrical networks (Fouss et al. 2007). This property only holds in the case of an undirected graph.

[13] Recall that the target node is always transformed into an absorbing and killing node.

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

$$\Delta_{st}^{\text{TsRSP}} = \langle c \rangle_{st} + \langle c \rangle_{ts} \tag{20}$$

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

5 Experimental comparisons

In this section, we evaluate our methods by comparing them with other state-of-the-art dissimilarity measures on node clustering and semi-supervised classification tasks.[14] Notice that our goal in this work is not to propose new node clustering or semi-supervised classification algorithms outperforming the state-of-the-art techniques. Our aim is instead to investigate if the Tsallis-based RSP model is able to capture the community structure of networks in an accurate way compared to other, already known, dissimilarity measures between nodes.

5.1 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. The global experimental setup, for the clustering task, is similar to the one used in Sommer et al. (2016) and Courtain et al. (2020) and, for the semi-supervised classification task, to Françoisse et al. (2017) and Guex et al. (2020). The description that follows is therefore largely inspired by these works.

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 Eq. (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 and four similarity matrices between nodes (kernels on a graph + the modularity matrix) will be used as baseline methods to assess our methods.

5.1.1 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 $\theta = 1/T$. These methods, already presented in

[14] Note that all results were obtained with Matlab (version R2019) running on an Intel Xeon with 2×8 core processors of 3.6 GHz and 128 GB of RAM.

Sect. 2, have been shown to perform well in both node clustering (Sommer et al. 2016) and semi-supervised classification tasks (Françoisse et al. 2017).
- The **Shortest Path** distance (SP) between two nodes i and j, corresponding to the cost associated with the least-cost path between them, computed from $\mathbf{C}$.
- The **Logarithmic Forest** distance (LF), defining a family of distances that interpolates, up to a scaling factor, between the shortest path distance and the resistance distance (Klein and Randic 1993), depending on a parameter α. It was introduced by Chebotarev (2011) based on the matrix-forest theorem (Chebotarev and Shamis 1997).

5.1.2 Baseline kernels

- The **Neumann** kernel (Schölkopf and Smola 2002) (Katz), initially proposed by Katz (1953) as a method of computing degrees of influence between nodes, and defined as $\mathbf{K} = (\mathbf{I} - \alpha\mathbf{A})^{-1} - \mathbf{I}$ for an undirected graph. The α parameter has to be positive and smaller than the inverse of the spectral radium of $\mathbf{A}$, $\rho(\mathbf{A}) = \max_i(|\lambda_i|)$.
- The **Logarithmic Communicability** kernel (lCom) proposed by Ivashkin et al. (2016) as the logarithmic version of the exponential diffusion kernel (Kondor and Lafferty 2002), also known as the communicability measure (Estrada and Hatano 2008), $\mathbf{K} = \ln(\exp(t\mathbf{A}))$, $t > 0$, where expm is the matrix exponential.
- The **Sigmoid Commute Time** kernel (SCT), proposed by Yen et al. (2007). This is obtained by applying a sigmoid transform (Schölkopf and Smola 2002), whose sharpness is controlled by a parameter α, on the commute time kernel (Fouss et al. 2007).

In addition, the **Modularity matrix Q** (which is not a valid kernel) is used as the last baseline (Modularity). The matrix is computed by $\mathbf{Q} = \mathbf{A} - \frac{\mathbf{d}\mathbf{d}^{\mathrm{T}}}{\mathrm{vol}}$ where $\mathbf{d}$ contains the node degrees and the constant vol is the volume of the graph (see, e.g., Newman 2018).

The experiments do not pretend to be comprehensive: our primary goal is to compare the introduced Tsallis-based dissimilarities to the free energy distance (FE) which obtained state-of-the-art results on similar tasks. The Katz, Modularity and SP methods are also chosen because they are standard, widely used, (dis)-similarity matrices. The other methods (RSP, LF, lCom, SCT) were also selected because of their good performances obtained in previous experimental comparisons (Courtain et al. 2020; Françoisse et al. 2017; Ivashkin et al. 2016; Sommer et al. 2016, 2017; Yen et al. 2007, 2008).

5.1.3 Datasets

A collection of 22 datasets, representing labeled undirected network, is investigated for the experimental comparisons of the dissimilarity measures. The collection includes Zachary's karate club (Zachary 1977), the Dolphin datasets (Lusseau 2003; Lusseau

Table 1 Datasets (networks) used for the experiments

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	

et al. 2003), the Football dataset (Girvan and Newman 2002), the Political books,[15] three LFR benchmarks (Lancichinetti et al. 2008), the WebKB datasets (Macskassy and Provost 2007), the IMDB dataset (Macskassy and Provost 2007), and nine Newsgroup datasets (Lang 1995; Yen et al. 2009).

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

5.2 Node clustering experiment

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

[15] Collected by V. Krebs and labeled by M.E. Newman, this dataset is not published, but is available for download at http://www-personal.umich.edu/~mejn/netdata/.

5.2.1 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)** (Fred and Jain 2003; Strehl and Ghosh 2002) between two partitions $\mathcal{U}$ and $\mathcal{V}$ is computed by dividing their mutual information (Cover and Thomas 2006) by the mean of their respective entropy (see, for example, Manning et al. 2008).
- The **Adjusted Rand Index (ARI)** (Hubert and Arabie 1985) is an extension of the Rand Index (Rand 1971), which measures the degree of overlap between two partitions. The ARI is adjusted to have an expected value of 0 for two random partitions, which is not the case for the original Rand Index.

5.2.2 Experimental methodology

The experiment follows a global methodology similar to Courtain et al. (2020), inspired by Sommer et al. (2016) (see this work for more details), that relies on a kernel k-means (see, e.g., Fouss et al. 2016; Yen et al. 2009). For the methods that provide a dissimilarity matrix $\mathbf{D}$, a kernel $\mathbf{K}$ is computed from $\mathbf{D}$ using classical multidimensional scaling (Borg and Groenen 1997). In the process, any negative eigenvalue is set to zero to ensure that $\mathbf{K}$ is positive semi-definite. A kernel k-means (see, e.g., Fouss et al. 2016; Yen et al. 2009) is then executed 30 times on $\mathbf{K}$. Throughout these 30 trials, the partition maximizing the modularity [which is an unsupervised measure of the quality of a partition of the nodes (Newman 2018)] is kept. The NMI and ARI are then computed for this partition.

This process is repeated 30 times (for a total of 900 runs of the k-means) for a given method (kernel or dissimilarity matrix), with a given value of its parameter (for instance, θ in the case of methods based on RSP), on a specific dataset. The NMI and ARI are averaged across these 30 repetitions.

The parameters that are tuned (one parameter per method) are the θ for the FE and the RSP, in both standard (FE and RSP) and Tsallis versions (FETsallis and RSPTsallis), the α for the LF, the α for Katz, the t for the lCom, and finally the α 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 θ. The values tested for these parameters are listed in Table 2. Recall that the parameters of the algorithms are tuned using the modularity as metric (Sommer et al. 2017).

The different methods are assessed globally across all datasets using the same method as the one used by Sommer et al. (2016), based on a non-parametric Friedman-Nemenyi test (Demšar 2006). In addition, a Wilcoxon signed-rank test (Wilcoxon 1945) is performed pairwise to measure the significance (at level $\alpha = 0.05$) of the differences observed in the algorithms' performance.

Table 2 Parameter range for the investigated methods

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	$\alpha = (0.05,\ 0.10, \ldots,\ 0.95) \times (\rho(\mathbf{A}))^{-1}$
lCom	$t = (0.01,\ 0.02,\ 0.05,\ 0.1,\ 0.2,\ 0.5,\ 1,\ 2,\ 5,\ 10)$
SCT	$\alpha = (5,\ 10,\ 15,\ \ldots,\ 50)$

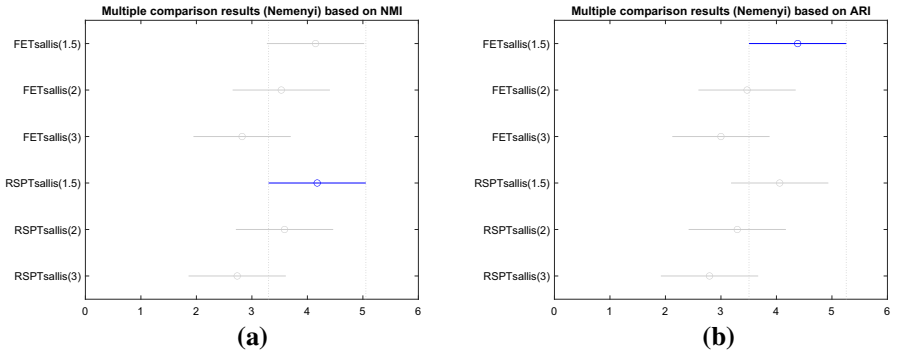

Fig. 3 Clustering experiment: analysis of the impact of the Tsallis divergence parameter r. Mean ranks and 95% Nemenyi confidence intervals for the FETsallis and RSPTsallis across the 17 datasets, according to (a) the NMI and (b) the ARI performance measures. Two methods are considered as significantly different if their confidence intervals do not overlap. The horizontal axis unit is the average rank of the methods. The higher the rank, the better the method; the best method is highlighted

5.2.3 Parameter analysis

Concerning 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 than the one based on the KL divergence.

A first preliminary experiment, following the procedure introduced in the previous subsection, shows that the parameter $r = 1.5$ for the Tsallis regularization seems to yield the best result out of the three values of r that were investigated, for both the FE and RSP, as shown in Fig. 3. It is, however, noteworthy that these differences are not significant according to the Wilcoxon signed-rank test (not shown for conciseness).

The next subsection therefore uses the value $r = 1.5$ for the FETsallis and the RSPTsallis for readability purposes.

5.2.4 Experimental results

The results of the experimental comparison with the baselines are displayed in Fig. 4. It can be seen that the Free Energy distance is the best method overall by a slight

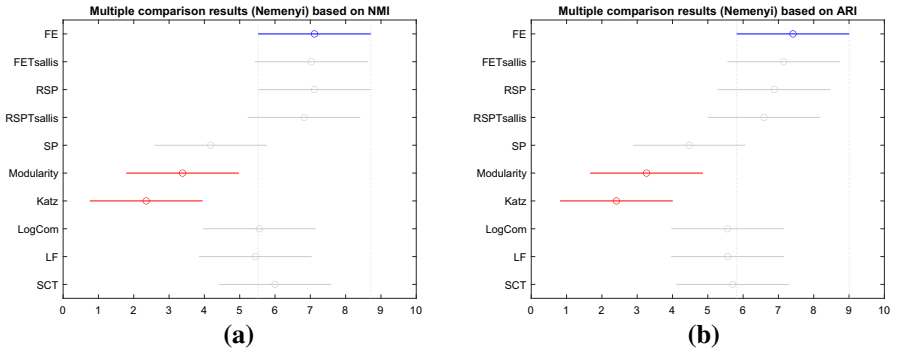

Fig. 4 Clustering experiment: overall comparison. Mean ranks and 95% Nemenyi confidence intervals for the 10 methods across the 17 datasets, according to (a) the NMI and (b) the ARI performance measures. Two methods are considered as significantly different if their confidence intervals do not overlap. The horizontal axis unit is the average rank of the methods. The higher the rank, the better the method; the best method is highlighted

margin and that, broadly speaking, the two versions of the FE and the RSP outperform the baselines. More precisely, the difference between these four methods and the Modularity matrix and Katz are significant. However, the difference in ranks with the LogCom, the LF, the SCT, and the SP is not significant according to the Friedman-Nemenyi test, which is rather conservative when the number of compared methods is large. This analysis holds for both the NMI and the ARI.

The pairwise Wilcoxon signed-rank tests also confirm no significant ($\alpha = 0.05$) difference in terms of ranks between the FE, the FETsallis, the RSP, and the RSPTsallis, for both the NMI and the ARI.

For the rest of the baselines, the Wilcoxon signed-rank test concludes that the differences in the performance with the FE and the FETsallis are all significant regarding the ARI. For the RSP and the RSPTsallis, the difference with the SP, the Modularity, and the Katz is significant, whereas the other differences are not. For the NMI, the conclusions for the Wilcoxon signed-rank test are globally the same with a few exepctions. Notably, the difference with the SCT is not significant for the FE, FETsallis, RSP, and RSPTsallis.

To sum up the results from the clustering task on the investigated datasets, the Tsallis regularized methods significantly outperform a majority of the baselines, and yields competitive results with respect to methods that have been shown to perform well in a context of kernel k-means clustering, notably the FE (Sommer et al. 2016).

5.3 Semi-supervised classification experiment

We now turn to the semi-supervised classification experiment.

5.3.1 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 that were hidden.

5.3.2 Experimental methodology

We followed the same experimental methodology as in Françoisse et al. (2017) and Guex et al. (2020). More precisely, this graph-based semi-supervised classification methodology consists in extracting the five[16] dominant eigenvectors of the kernel matrix (possibly derived from the dissimilarity matrix by classical multidimensional scaling), in order to use them as node features in a linear support vector machine (SVM).[17] The objective is therefore to evaluate if the different measures are able to capture the relevant information present in the graph structure in just a few dimensions. In other words, we want to assess if the measures are able to perform efficient feature extraction from the graph. Note that this setting is inspired by the work of Zhang and Mao (2008a, b) as well as Tang and Liu (2009a, b, 2010), 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 an SVM, to categorize the nodes.

All the methods are tested by using a standard 5×5 nested cross-validation methodology. Each external cross-validation contains five folds, and methods are tested with a labeling rate of 20%. To tune parameters (see Table 2 for values; one parameter per method is tuned), an internal five-fold cross-validation on the training fold is performed with a labeling rate of 80%. The whole cross-validation procedure is repeated five 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.

Concerning the parameter r of the Tsallis regularization, as for clustering, we tested three values $r = \{1.5, 2, 3\}$ in a preliminary analysis and kept only the overall best-performing value (the r parameter is not tuned on the datasets). For the SVM, the margin parameter is tuned on the set of values $c = \{10^{-2}, 10^{-1}, 1, 10, 100\}$.

5.3.3 Parameter analysis

As in the clustering experiment, we start our analysis by comparing the performances of the Tsallis regularized methods, for the three investigated values of parameter r—see Table 3 for the detailed results. Then, a Friedman-Nemenyi statistical test comparing the variants of the Tsallis FE and RSP is performed with a confidence level of 95% ($\alpha = 0.05$) (Demšar 2006). From the results presented in Fig. 5, we can observe that

[16] We arbitrarily report the results for five dimensions but also performed experiments with more dimensions with similar conclusions.

[17] We use the LIBLINEAR library Fan et al. (2008) with the options '-s 1' and '-B 1'.

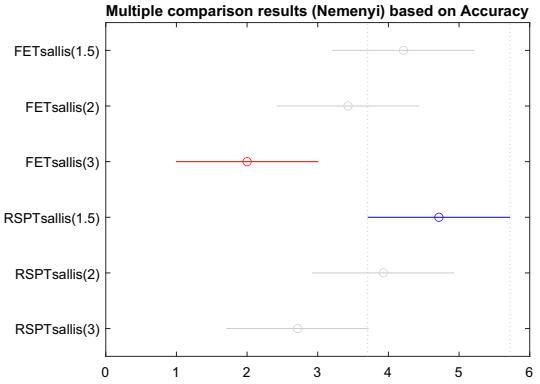

Fig. 5 Semi-supervised experiment: analysis of the impact of the Tsallis divergence parameter r. Mean ranks and 95% Nemenyi confidence intervals for the RSPTsallis and the FETsallis (see Table 3) across the 14 datasets, based on classification accuracy. Two methods are considered as significantly different if their confidence intervals do not overlap. The horizontal axis unit is the average rank of the methods. The higher the rank, the better the method. The best method overall (RSPTsallis(1.5)) is highlighted

the parameter $r = 1.5$ has the largest average rank for both the FETsallis and the RSPTsallis. Nevertheless, the Friedman-Nemenyi test does not reveal a significant difference between the results of the parameter values. Contrary to the results of the clustering experiment, the Wilcoxon tests (not shown for conciseness) highlight that the RSPTsallis with a parameter $r = 1.5$ performs significantly better than the RSPTsallis with a parameter $r = 2$ or $r = 3$. As regards the FETsallis, the Wilcoxon tests indicate that the FETsallis using a parameter $r = 1.5$ or $r = 2$ obtains significantly better results than the FETsallis with a parameter $r = 3$. From these findings and to avoid redundancy, we used only the parameter value $r = 1.5$ for the FETsallis and the RSPTsallis in the comparison with the baselines presented in the next subsection.

5.3.4 Experimental results

The results of this semi-supervised classification experiment are reported in Table 3. The highest accuracy is highlighted in boldface for each dataset. From Table 3, it can be seen that 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 Friedman-Nemenyi statistical test and pairwise Wilcoxon signed-rank tests at a confidence level of 95% ($\alpha = 0.05$) (Demšar 2006).

The results of the Friedman-Nemenyi test are shown in Fig. 6. Recall that, in this comparison, the parameter r of the Tsallis FE and RSP has been fixed to the value $r = 1.5$, following the preliminary parameter analysis. From this figure, it can be observed that the RSPTsallis is the method with the highest average rank overall. The test emphasizes that the RSPTsallis performs significantly better than the SP, the Modularity, and the Katz. Furthermore, the figure also highlights that the FE, the FETsallis, and the LogCom outperform the SP and the Katz. Finally, we observe that the RSP and the SCT provide results significantly superior to the performances of Katz.

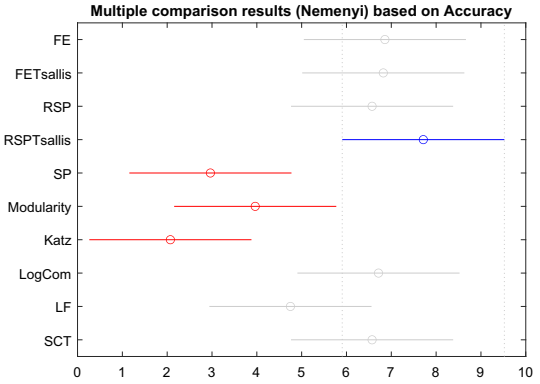

Fig. 6 Semi-supervised experiment: overall comparison. Mean ranks and 95% Nemenyi confidence intervals for the 10 methods (see Table 3) across the 14 datasets, based on classification accuracy. Two methods are considered as significantly different if their confidence intervals do not overlap. The horizontal axis unit is the average rank of the methods. The higher the rank, the better the method. The best method overall (RSPTsallis) is highlighted

Moreover, the Wilcoxon signed-rank tests show that the RSPTsallis performs significantly better than the LF. The tests also highlight that the FE, the FETsallis, and the LogCom obtain significantly better results than the Modularity. As regards the RSP, we notice that it outperforms the SP, the Modularity, and the LF. The LF, for its part, provides results significantly superior to the SP and the Katz. Finally, the tests indicate that the SCT performs significantly better than the SP and the Modularity. Nevertheless, no significant difference is observed, in terms of ranks, between the KL and the Tsallis divergence regularization for either the FE or the RSP.

Globally, this confirms that the Tsallis RSP and FE dissimilarity measures, and especially the Tsallis RSP, are able to capture the community structure of the graph in a competitive way, at least for the investigated datasets and methods. In addition, the LogCom kernel also provides highly competitive results, notably on the semi-supervised classification task where this method performs best on several datasets of different kind.

6 Conclusion and future work

This paper shows 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 (observed on sparse graphs) 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 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

Table 3 Classification accuracies (%) for the various classification methods obtained on the different datasets

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	**93.13**
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	**93.57**	93.07
Newsgroup_3_3	**93.66**	93.35	92.29	91.83	92.91	93.53	93.03
Newsgroup_5_1	**88.70**	88.54	88.24	87.62	88.09	88.22	88.03
Newsgroup_5_2	82.32	**82.55**	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	**83.42**
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	**78.53**	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	**97.14**
Newsgroup_2_2	92.79	90.36	91.22	90.38	91.18	90.36	91.14
Newsgroup_2_3	96.32	**96.78**	95.78	93.21	95.43	95.54	95.79
Newsgroup_3_1	92.54	93.01	93.02	91.00	**94.00**	92.98	93.99

Table 3 continued

Classif. method → Dataset ↓	RSPTsallis(3)	SP	Modularity	Katz	LogCom	LF	SCT
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	**59.16**	58.35	57.30
WebKB-Texas	72.93	61.20	73.01	67.89	**75.45**	74.18	74.27
WebKB-Washington	64.69	52.35	62.52	64.56	**69.92**	67.47	67.66
WebKB-Wisconsin	69.64	62.46	73.42	73.61	**74.99**	74.48	72.77

For each dataset and method, the final accuracy is obtained by averaging over five repetitions of a standard cross-validation procedure. Each repetition consists of a nested cross-validation with five external folds (test sets, for validation) on which the accuracy of the classifier is averaged, and five internal folds (for parameter tuning). The best performing method is highlighted in boldface for each dataset

of the network for tackling data mining and machine learning tasks. A nice property is that they interpolate between the shortest path distance (when $\theta \to \infty$) and the commute cost distance (proportional to the resistance distance, $\theta \to 0^+$). Another interesting property is the fact that, as the standard randomized shortest paths, in the computation of the dissimilarity, they take into account not only the proximity but also the degree of inter-connectivity (direct and indirect).

Indeed, it is well known that the standard shortest path distance and the resistance distance, while very useful in many contexts, have 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., Fouss et al. 2016)], while the resistance distance sometimes provides useless results when dealing with large graphs [the so-called lost in space effect (von Luxburg et al. 2010, 2014)]. Another drawback of the shortest path distance is that it usually provides a large number of ties when comparing distances, especially on unweighted and undirected graphs. Moreover, it has been shown recently (Hashimoto et al. 2015) that the "logarithmic Laplace transformed hitting time" (closely related to the FE distance based on Kullback–Leibler regularization) avoids to a certain extent 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 data mining 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 to large graphs.

Future 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 by Duchi et al. (2008) to compute the orthogonal projection on the unit simplex, based on a modification of the randomized median finding procedure (Blum et al. 1973; Cormen et al. 2009). The link between the orthogonal projection on the unit simplex and our formulation (12) should also be studied.

Another interesting question, raised by a reviewer, concerns why certain methods perform well on some graphs and what this says about those graphs. A potential response would be to explore the properties of a large number of graphs that have different structures[18] and then to investigate which method works better/worse, depending on these properties. Indeed, each similarity/dissimilarity measure captures different structural properties of the graph, and it would be useful to study their impact on the results. This is left for future work.

Yet another interesting contribution would be to apply the Tsallis divergence regularization to the solution of Markov decision problems to induce sparse policies, thereby extending the work of Lee et al. (2018b) in such a way that the solution interpolates between a random walk and an optimal, least-cost, behavior. We also plan to

[18] There is a vast literature on this subject; see for instance Newman (2018) as a starting point.

use the Tsallis divergence for the design of algorithms solving the sparse randomized optimal transport on a graph problem with multiple sources and multiple destinations, again building on previous work (Guex 2016; Guex et al. 2019). It would also be interesting to investigate a Renyi divergence regularization instead of the Kullback–Leibler or the Tsallis one.

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 Francois Bavaud (Université de Lausanne) and Dr Amin Mantrach for their interesting remarks and discussions. Finally, we are also grateful to the anonymous reviewers whose suggestions have helped us significantly to improve the manuscript.

Appendices

These appendices discuss the alternative form of the randomized shortest paths, the convexity of the objective function with Tsallis divergence regularization, and the algorithms for computing the spmin function appearing in Eq. (13), returning transition probabilities $\mathbf{p}_i$ associated with a node i. To the best of our knowledge, these algorithms were first studied by Kanzawa (2013) and Miyamoto and Umayahara (1998) (although we have the feeling that they had been investigated before, maybe in the context of the closely related projection on the probability simplex, as discussed in the related work section). This appendix provides a reformulation of the relevant material contained in these papers (Appendices C.1 and 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, Appendix 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 (Akamatsu 1997; García-Díez et al. 2011a; Saerens et al. 2009) 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 $\eta\big((i,j) \in \wp\big)$, defined as the number of times edge (i,j) is visited along path $\wp$. Because the probability of a path can be expressed as a product of transition probabilities [see the discussion after Eq. (4)], the path-based quantities $\log(\mathrm{P}(\wp)/\tilde{\pi}(\wp))$ and $\tilde{c}(\wp)$ can be expressed as

$$\begin{cases} \log \dfrac{\mathrm{P}(\wp)}{\tilde{\pi}(\wp)} = \sum_{(i,j)\in\mathcal{E}} \eta\big((i,j)\in\wp\big) \log \dfrac{p_{ij}}{p_{ij}^{\mathrm{ref}}} \\ \tilde{c}(\wp) = \sum_{(i,j)\in\mathcal{E}} \eta\big((i,j)\in\wp\big)\, c_{ij} \end{cases} \tag{A.1}$$

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

$$\begin{aligned} \phi_{st}^{\mathrm{KL}}(\mathrm{P}) &= \sum_{\wp\in\mathcal{P}_{st}} \mathrm{P}(\wp)\tilde{c}(\wp) + T \sum_{\wp\in\mathcal{P}_{st}} \mathrm{P}(\wp) \log\left(\frac{\mathrm{P}(\wp)}{\tilde{\pi}(\wp)}\right) \\ &= \sum_{\wp\in\mathcal{P}_{st}} \mathrm{P}(\wp) \sum_{(i,j)\in\mathcal{E}} \eta\big((i,j)\in\wp\big)\, c_{ij} + T \sum_{\wp\in\mathcal{P}_{st}} \mathrm{P}(\wp) \sum_{(i,j)\in\mathcal{E}} \eta\big((i,j)\in\wp\big) \\ &\quad \log \frac{p_{ij}}{p_{ij}^{\mathrm{ref}}} \\ &= \sum_{(i,j)\in\mathcal{E}} \bar{n}_{ij}\, c_{ij} + T \sum_{(i,j)\in\mathcal{E}} \bar{n}_{ij} \log \frac{p_{ij}}{p_{ij}^{\mathrm{ref}}} \\ &= \sum_{(i,j)\in\mathcal{E}} \bar{n}_i\, p_{ij} \left(c_{ij} + T \log \frac{p_{ij}}{p_{ij}^{\mathrm{ref}}} \right) \end{aligned} \tag{A.2}$$

where $\bar{n}_{ij} \triangleq \sum_{\wp\in\mathcal{P}_{st}} \mathrm{P}(\wp)\, \eta\big((i,j)\in\wp\big)$ is the expected number of passages (the flow) through edge (i,j) and we use $\bar{n}_{ij} = \bar{n}_i\, p_{ij}$ with $\bar{n}_i = \bar{n}_{i\bullet} = \sum_{j\in\mathcal{S}ucc(i)} \bar{n}_{ij}$ to denote 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 Bavaud and Guex (2012) and Guex and Bavaud (2015) based on edge flows and flow conservation.

The objective function (A.2) should be minimized with respect to the (local) policy, that is, the set of transition probabilities p_{ij} associated with edges. It is shown in Akamatsu (1997) that this objective function is strictly convex with respect to edge flows, $\bar{n}_{ij} = \bar{n}_i\, p_{ij}$ with $\bar{n}_i = \sum_{j\in\mathcal{S}ucc(i)} \bar{n}_{ij}$. Indeed, the objective function, which is expressed in Eq. (A.2) in terms of transition probabilities and number of visits to nodes, can also be expressed in terms of edge flows only, or in terms of transition probabilities only. Therefore, we conjecture that the objective function is also convex with respect to the transition probabilities inside the probability simplex. Indeed, because the correspondence between edge flows and transition probabilities is strictly increasing, differentiable, and one-to-one for a unit input flow (which is indeed the case in the RSP model), any minimum of Eq.(A.2) with respect to the edge flows should also be a minimum with respect to the corresponding transition probabilities, and vice-versa. Thus, because the objective function is convex with respect to the $\bar{n}_{ij}$ and the domain is convex, it should be a global minimum. However, this heuristic argument is a simple conjecture at this point and should be verified formally.

Interestingly, this also shows that the path-based formalism of Eq. (2) is equivalent to minimizing the local cost plus KL divergence, $c_{ij} + T \log(p_{ij}/p_{ij}^{\mathrm{ref}})$, 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 review of related work (Sect. 1.2), the randomized shortest paths framework is equivalent to some of these models (see, e.g., Busic and Meyn 2018; Fox et al. 2001; Kappen et al. 2012; Rubin et al. 2012; Theodorou et al. 2013; Theodorou and Todorov 2012), as initiated by Todorov (2007, 2008).

A.2 Computation of the optimal policy

In this subsection, an algorithm for computing the optimal transition probabilities (the policy) is developed. It aims to minimize the objective function (A.2) 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., Bryson and Ho 1975; Luenberger 1979). After *renumbering the nodes* in such a way that node 1 is the source node and node n the target node[19] for convenience, this leads to the following Lagrange function only including the equality constraints,

$$\begin{aligned}\mathscr{L}(\mathbf{P}, \bar{\mathbf{n}}; \boldsymbol{\mu}, \boldsymbol{\lambda}_{\mathrm{KL}}) = &\sum_{i\in\mathcal{V}\setminus n} \bar{n}_i \sum_{j\in\mathcal{S}ucc(i)} p_{ij}\left(c_{ij} + T\log\frac{p_{ij}}{p_{ij}^{\mathrm{ref}}}\right) \\ &+ \sum_{i\in\mathcal{V}\setminus n} \mu_i\left(1 - \sum_{j\in\mathcal{S}ucc(i)} p_{ij}\right) \\ &+ \sum_{j\in\mathcal{V}} \lambda_j^{\mathrm{KL}}\left(\sum_{i\in\mathcal{P}red(j)} \bar{n}_i p_{ij} + \delta_{1j} - \bar{n}_j\right) \end{aligned} \tag{A.3}$$

where $\mathcal{S}ucc(i)$ is the set of successor nodes[20] of node i and $\mathcal{P}red(j)$ is the set of predecessor nodes of node j. The quantities μ_i, λ_i^{KL} are Lagrange parameters dealing with equality 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., Norris 1997; Taylor and Karlin 1998). 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 (Beck 2014; Culioli 2012; Griva et al. 2008; Minoux 1986). 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 λ_i^{KL}. We observe that they only depend on the Lagrange parameters, which are then computed by maximizing the dual, a common optimization procedure closely related to the Arrow-Hurwicz-Uzawa algorithm (Arrow et al. 1958). The two steps are iterated until convergence, which is guaranteed because each sub-problem reaches its optimum uniquely (Bertsekas

[19] It is assumed that the source node is different from the target node.

[20] Recall that target node n is killing and absorbing, and thus has no successor. It is therefore not the predecessor of any node.

1999). In practice, we observed in our experiments that the duality gap is always zero, showing that a global minimum is reached.

A.2.1 Computation of the transition probabilities

For the estimation of the transition probabilities, there is no need to introduce non-negativity constraints, because KL divergence regularization ensures that they satisfy the constraint (Kapur 1989). Taking the partial derivative of the Lagrange function (A.3) with respect to the transition probabilities 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

$$p_{ij} = \frac{p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \lambda_j^{\text{KL}})]}{\sum_{k \in \mathcal{S}ucc(i)} p_{ik}^{\text{ref}} \exp[-\theta(c_{ik} + \lambda_k^{\text{KL}})]} \quad \text{for all } (i, j) \in \mathcal{E} \tag{A.4}$$

which only depends on the Lagrange parameters λ_i^{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.

A.2.2 Computation of the Lagrange parameters

Computing the λ_i^{KL} aims to solve the dual problem. Indeed, by defining respectively the expected cost and the KL divergence per node, $\tilde{c}_i = \sum_{j \in \mathcal{S}ucc(i)} p_{ij} c_{ij}$ and $\bar{h}_i^{\text{KL}} = \sum_{j \in \mathcal{S}ucc(i)} p_{ij} \log(p_{ij}/p_{ij}^{\text{ref}})$ for $i \neq n$, together with $\bar{c}_n = 0$ and $\bar{h}_n^{\text{KL}} = 0$ for the target node, the problem of computing the expected number of visits $\bar{n}_i$ when transition probabilities are fixed can be reformulated from Eq. (A.3) as a linear programming problem[21]: minimize $(\bar{\mathbf{c}} + T\bar{\mathbf{h}}_{\text{KL}})^{\text{T}}\bar{\mathbf{n}}$ with respect to $\bar{\mathbf{n}}$ subject to the constraints $(\mathbf{I} - \mathbf{P})^{\text{T}}\bar{\mathbf{n}} = \mathbf{e}_1$ and $\bar{\mathbf{n}} \geq \mathbf{0}$.

However, because we need the vector of Lagrange parameters $\boldsymbol{\lambda}_{\text{KL}}$ in order to compute the transition probabilities [see Eq. (A.4)], we are more interested in the dual problem (Griva et al. 2008, pages 184 and 533),[22]

$$\left| \begin{array}{l} \underset{\boldsymbol{\lambda}_{\text{KL}}}{\text{maximize}} \ \mathbf{e}_1^{\text{T}}\boldsymbol{\lambda}_{\text{KL}} \\ \text{subject to} \ (\mathbf{I} - \mathbf{P})\boldsymbol{\lambda}_{\text{KL}} = \bar{\mathbf{c}} + T\bar{\mathbf{h}}_{\text{KL}} \\ \qquad\qquad \boldsymbol{\lambda}_{\text{KL}} \geq \mathbf{0} \end{array} \right. \tag{A.5}$$

[21] Alternatively and maybe more directly, the resulting Eq. (A.7) for computing the Lagrange parameters can also be obtained by stating the Karush–Kuhn–Tucker necessary conditions; namely by setting the partial derivative of the Lagrange function (10) with respect to the n_i to zero Bryson and Ho (1975); Griva et al. (2008).

[22] Note that the equation in Griva et al. (2008) also holds for equality constraints, providing $(\mathbf{I}-\mathbf{P})^{\text{T}}\bar{\mathbf{n}} = \mathbf{e}_1$, which is the case here.

where $\mathbf{e}_1$ is a column vector full of zeros, except in position 1 containing a 1. Because the elements of $(\mathbf{I}-\mathbf{P})^{-1} = \mathbf{I}+\mathbf{P}+\mathbf{P}^2+\cdots$, $\bar{\mathbf{c}}$ and $\bar{\mathbf{h}}_{\mathrm{KL}}$ 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

$$(\mathbf{I}-\mathbf{P})\boldsymbol{\lambda}_{\mathrm{KL}} = \bar{\mathbf{c}} + T\bar{\mathbf{h}}_{\mathrm{KL}} \tag{A.6}$$

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

$$\lambda_i^{\mathrm{KL}} = \sum_{j\in\mathcal{S}ucc(i)} p_{ij}\left(c_{ij}+\lambda_j^{\mathrm{KL}}+T\log\frac{p_{ij}}{p_{ij}^{\mathrm{ref}}}\right) \tag{A.7}$$

Notice that this implies $\lambda_n^{\mathrm{KL}} = 0$ for the target node n. The procedure aims to iterate Eqs. (A.4) and (A.6) until convergence. The Lagrange parameters have an interesting interpretation, as explained in the next subsection.

A.2.3 Interpretation of the Lagrange parameters

Let us now give an interpretation to the Lagrange parameters $\boldsymbol{\lambda}_{\mathrm{KL}}$ at the optimum. From Eq. (A.4), we directly obtain

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

By injecting this expression in Eq. (A.7), we obtain

$$\lambda_i^{\mathrm{KL}} = -\tfrac{1}{\theta}\log\sum_{j\in\mathcal{S}ucc(i)} p_{ij}^{\mathrm{ref}}\exp[-\theta(c_{ij}+\lambda_j^{\mathrm{KL}})] \tag{A.8}$$

which is nothing other than the Bellman–Ford-like recurrence formula[23] [Françoisse et al. 2017, Eq. (34)][24] for computing the directed *free energy distance* $\boldsymbol{\phi}^{\mathrm{KL}}$ defined in Eq. (2). Therefore $\boldsymbol{\lambda}^{\mathrm{KL}} = \boldsymbol{\phi}^{\mathrm{KL}}$ 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 Eq. (3)] in the free energy objective function provided by Eq. (2) (Kivimäki et al. 2014). This quantity has many interesting properties: (1) it plays the role of a potential at the continuous space-time limit (García-Díez et al. 2011b), (2) it defines a distance measure between nodes when symmetrized (Françoisse et al. 2017; Kivimäki et al. 2014), (3) it interpolates between the least-cost and the

[23] The min operator in the Bellman–Ford expression is replaced by a softmin, or log-sum-exp, operator (Bloem and Bambos 2014; Françoisse et al. 2013, 2017).

[24] As it plays the role of a potential, the FE distance was called the potential distance in Françoisse et al. (2013, 2017).

commute cost distance, (4) it can be interpreted as minus T times the log-likelihood of surviving during a particular killed random walk (Françoisse et al. 2017), and (5) it performed consistently well in a number of data mining tasks (Françoisse et al. 2017; Sommer et al. 2016, 2017), among others.

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 Eq. (9). To this end, we will compute the Hessian matrix with respect to the flows $\bar{n}_{ij}$ and verify if the corresponding quadratic form is positive semi-definite. Using $\bar{n}_{ij} = \bar{n}_i\, p_{ij}$ with $\bar{n}_i = \bar{n}_{i\bullet} = \sum_{j\in\mathcal{S}ucc(i)} \bar{n}_{ij}$, we first reformulate this objective function in terms of edge flows $\bar{n}_{ij}$,

$$\begin{aligned}\phi_{st}^{\mathrm{Ts}} &= \sum_{i\in\mathcal{V}} \bar{n}_{i\bullet} \sum_{j\in\mathcal{S}ucc(i)} p_{ij}\left(c_{ij} + \tfrac{T}{r-1}\left(\left(\frac{p_{ij}}{p_{ij}^{\mathrm{ref}}}\right)^{r-1} - 1\right)\right)\\ &= \sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{S}ucc(i)} \left(c_{ij}\bar{n}_{ij} + \tfrac{T}{r-1}\frac{(\bar{n}_{ij})^r}{(p_{ij}^{\mathrm{ref}}\bar{n}_{i\bullet})^{r-1}} - \tfrac{T}{r-1}\bar{n}_{ij}\right)\end{aligned} \tag{B.1}$$

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}^{ref} for simplicity),

$$f(\mathbf{N}) = \sum_{k\in\mathcal{V}} \sum_{l\in\mathcal{S}ucc(k)} \frac{(\bar{n}_{kl})^r}{(\bar{n}_{k\bullet})^{r-1}} \tag{B.2}$$

The first-order partial derivatives are

$$\frac{\partial f}{\partial \bar{n}_{ij}} = r\, p_{ij}^{r-1} - (r-1)\textstyle\sum_{j'\in\mathcal{S}ucc(i)} p_{ij'}^r \tag{B.3}$$

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

$$h_{(i,j)(k,l)} = \frac{\partial^2 f}{\partial \bar{n}_{kl}\partial \bar{n}_{ij}} = r(r-1)\frac{\delta_{ik}}{\bar{n}_{i\bullet}}\left(\delta_{jl}\, p_{ij}^{r-2} - (p_{ij}^{r-1} + p_{il}^{r-1}) + \textstyle\sum_{j'\in\mathcal{S}ucc(i)} p_{ij'}^r\right) \tag{B.4}$$

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

$$\begin{aligned} Q &= \sum_{(i,j)\in\mathcal{E}} \sum_{(k,l)\in\mathcal{E}} x_{(i,j)} h_{(i,j)(k,l)} x_{(k,l)}\\ &= r(r-1)\sum_{i\in\mathcal{V}} \frac{1}{\bar{n}_{i\bullet}} \sum_{j,l\in\mathcal{S}ucc(i)} x_{(i,j)}\\ &\quad \underbrace{\left(\delta_{jl}\, p_{ij}^{r-2} - (p_{ij}^{r-1} + p_{il}^{r-1}) + \textstyle\sum_{j'\in\mathcal{S}ucc(i)} p_{ij'}^r\right)}_{\text{matrix } q_{jl}(i)} x_{(i,l)} \end{aligned} \tag{B.5}$$

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

$$f(\mathbf{N}) = \sum_{k\in\mathcal{V}} \sum_{l\in\mathcal{S}ucc(k)} (p_{kl}^{\mathrm{ref}})^{1-r} \frac{(\bar{n}_{kl})^r}{(\bar{n}_{k\bullet})^{r-1}} \tag{B.6}$$

and the Hessian is

$$h_{(i,j)(k,l)} = r(r-1)\frac{\delta_{ik}}{\bar{n}_{i\bullet}}\Big(\delta_{jl}(p_{ij}^{\mathrm{ref}})^{1-r}\, p_{ij}^{r-2} - \big((p_{ij}^{\mathrm{ref}})^{1-r}\, p_{ij}^{r-1} + (p_{il}^{\mathrm{ref}})^{1-r}\, p_{il}^{r-1}\big) + \textstyle\sum_{j'\in\mathcal{S}ucc(i)}(p_{ij'}^{\mathrm{ref}})^{1-r}\, p_{ij'}^{r}\Big) \tag{B.7}$$

Then, the quadratic form can be readily deduced

$$\begin{aligned} Q = r(r-1)\sum_{i\in\mathcal{V}}\frac{1}{\bar{n}_{i\bullet}} \sum_{j,l\in\mathcal{S}ucc(i)} x_{(i,j)} \\ \times \Big(\delta_{jl}(p_{ij}^{\mathrm{ref}})^{1-r}\, p_{ij}^{r-2} - \big((p_{ij}^{\mathrm{ref}})^{1-r}\, p_{ij}^{r-1} + (p_{il}^{\mathrm{ref}})^{1-r}\, p_{il}^{r-1}\big) \\ + \textstyle\sum_{j'\in\mathcal{S}ucc(i)}(p_{ij'}^{\mathrm{ref}})^{1-r}\, p_{ij'}^{r}\Big) x_{(i,l)} \end{aligned} \tag{B.8}$$

Because we did not find an easy way to prove formally the positive semi-definiteness of the matrices $\mathbf{Q}(i)$ with elements $q_{jl}(i) = \delta_{jl}(p_{ij}^{\mathrm{ref}})^{1-r}\, p_{ij}^{r-2} - \big((p_{ij}^{\mathrm{ref}})^{1-r}\, p_{ij}^{r-1} + (p_{il}^{\mathrm{ref}})^{1-r}\, p_{il}^{r-1}\big) + \sum_{j'\in\mathcal{S}ucc(i)}(p_{ij'}^{\mathrm{ref}})^{1-r}\, p_{ij'}^{r}$, we decided to test it numerically. More precisely, we generated 10^6 random instances of 20-dimensional probability vectors $\mathbf{p}_i$ and $\mathbf{p}_i^{\mathrm{ref}}$, 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 some evidence that the objective function is convex with respect to the flows n_{ij}, although it remains a conjecture at this state. Then, by using the heuristic argument presented at the end of Appendix A.1 of this appendix [after Eq. (A.2)], it should also be convex with respect to the transition probabilities. But this is still another conjecture that needs to be verified in a formal way.

C Minimizing expected cost plus Tsallis divergence

In this section, for completeness, we show how to solve the problem stated in Eq. (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 (Appendix C.1). Then, we extend the results to the more general $r > 1$ case (Appendix C.2). Finally, the most general case of $r > 1$ and a non-uniform reference probability distribution (and thus Tsallis divergence

regularization) of Eq. (12) is considered (C.3). Note that the Appendices C.1 and C.2 are based on the works of Kanzawa (2013) and Miyamoto and Umayahara (1998); it is simply a reformulation of their results.

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

This subsection derives the algorithm for finding a possibly sparse solution to the problem of minimizing an expected cost under quadratic constraints (or regularization), thus in the case where $r = 2$ and a uniform reference distribution (adapted from Kanzawa 2013; Miyamoto and Umayahara 1998). This is the simplest case; the problem stated in Eq. (12) then reduces to

$$\left| \begin{array}{l} \underset{\mathbf{p}}{\text{minimize}}\ \mathbf{c}^{\mathrm{T}}\mathbf{p} + T\,\|\mathbf{p}\|_2^2 \\ \text{subject to}\ \mathbf{e}^{\mathrm{T}}\mathbf{p} = 1 \\ \qquad\qquad \mathbf{p} \geq \mathbf{0} \end{array} \right. \tag{C.1}$$

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 the 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 with the outgoing edges of a transient node i, and $\mathbf{p}$ corresponds to its transition probabilities $\mathbf{p}_i$ [see Eq. (12)].

C.1.1 The Karush–Kuhn–Tucker conditions

By using the necessary Karush–Kuhn–Tucker conditions (Griva et al. 2008; Luenberger and Ye 2010; Rardin 1998) and introducing the Lagrange parameters μ [equality constraint in Eq. (C.1)] and $\boldsymbol{\lambda}$ [non-negativity constraints in Eq. (C.1)], we obtain, in addition to the sum-to-one and non-negativity constraints on $\mathbf{p}$,

$$\begin{cases} \mathbf{c} + 2T\mathbf{p} - \mu\mathbf{e} - \boldsymbol{\lambda} = \mathbf{0} \\ \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{p} = 0 \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \tag{C.2}$$

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: the two quantities cannot be both positive. We further denote by $\mathcal{Q}_+^*$ the set of nodes with a *strictly positive* p_i (to be found) and $|\mathcal{Q}_+^*|$ the number of such nodes.

Let us now consider the different cases. First, if $p_i = 0$ for element i, Eq. (C.2) tells us that $\lambda_i = c_i - \mu$. 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 $\lambda_i = 0$).

Now that we know what happens when $c_i < \mu$, let us investigate the situation where $c_i \geq \mu$, equivalent to $\mu - c_i \leq 0$. From Eq. (C.2), 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 $c_i \geq \mu$. Indeed, it suffices to consider the two cases $\lambda_i > 0$ and $\lambda_i = 0$, both leading to $p_i = 0$. The parameter μ is therefore a *threshold* on the cost 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 Eq. (C.2), $c_i + 2Tp_i - \mu = 0$, which provides $p_i = \frac{1}{2T}(\mu - c_i)$. Therefore, the p_i are of the following form

$$p_i = \tfrac{1}{2T}[\mu - c_i]_+ = \begin{cases} \frac{1}{2T}(\mu - c_i) & \text{when } \mu - c_i > 0 \\ 0 & \text{when } \mu - c_i \leq 0 \end{cases} \tag{C.3}$$

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, when the cost reaches the threshold μ. 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, $c_1 \leq c_2 \leq \cdots \leq c_m$. In that case, from Eq. (C.3), the sequence of $(p_i)_{i=1}^m$ is monotonic non-increasing for a fixed μ.

Equation (C.3) also tells us that the optimal set of nodes with a strictly positive p_i is given by $\mathcal{Q}_+^* = \{1, 2, \ldots, k^*\}$, for some index $k^* = |\mathcal{Q}_+^*|$ to be found. We now have to compute the Lagrange parameter μ as well as this threshold index k^*.

C.1.2 Computing the Lagrange parameter μ

By expressing the sum-to-one constraint on $\mathbf{p}$ based on Eq. (C.3) assuming that the optimal number of strictly positive elements $|\mathcal{Q}_+^*| = k^*$ is known, the Lagrange parameter can easily be computed, $\mu = \frac{1}{|\mathcal{Q}_+^*|}(\sum_{j\in\mathcal{Q}_+^*} c_j) + \frac{2T}{|\mathcal{Q}_+^*|}$. Further injecting this expression into Eq. (C.3), we obtain for $i \in \mathcal{Q}_+^*$ and thus $i \leq k^*$

$$p_i = \tfrac{1}{2T}\left(\tfrac{1}{|\mathcal{Q}_+^*|}\sum_{j\in\mathcal{Q}_+^*} c_j - c_i\right) + \frac{1}{|\mathcal{Q}_+^*|} = \tfrac{1}{2T}\left(\tfrac{1}{k^*}\underbrace{\sum_{j=1}^{k^*} c_j}_{=s(k^*)} - c_i\right) + \frac{1}{k^*} \tag{C.4}$$

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

$$p_i = \left[\tfrac{1}{2T}\left(\tfrac{1}{k^*}\sum_{j=1}^{k^*} c_j - c_i\right) + \frac{1}{k^*}\right]_+ \tag{C.5}$$

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

C.1.3 Computing the threshold index k^*

The idea now for finding k^* is to investigate sequentially a number of positive elements $k = 1, 2, \ldots$ 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 Eq. (C.3), 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

$$\sum_{i=1}^{k} p_i = \frac{1}{2T} \sum_{i=1}^{k} (\mu - c_i) = \frac{1}{2T} \left(k\mu - \sum_{i=1}^{k} c_i \right) \tag{C.6}$$

The optimal threshold μ 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[25] the L1 norm is equal to or exceeds 1, as proposed by Kanzawa (2013).

To this end, let us introduce the following auxiliary function from Eq. (C.6), corresponding to the L1 norm when considering an increasing sequence of length $k = 1, 2, \ldots$ and corresponding discrete values $\mu = c_k$ for μ,

$$L_1(k) = \frac{1}{2T} \left(k c_k - \underbrace{\sum_{i=1}^{k} c_i}_{s(k)} \right) = \frac{1}{2T} \big(k c_k - s(k) \big) \tag{C.7}$$

with $s(k) = \sum_{i=1}^{k} c_i$. Obviously, $L_1(1) = 0$. Moreover, the function $L_1(k)$ 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 μ 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 $L_1(k)$ sequentially and stopping when $L_1(k^* + 1) \geq 1$ becomes true. If this last condition is never reached, that is, $L_1(m) < 1$, this means that all the m elements of $\mathbf{p}$ are strictly positive and thus $k^* = m$.

C.1.4 Computing the optimal probability distribution

Once the number of positive elements k^* is computed, the probability mass $\mathbf{p}$ can be obtained from Eq. (C.4) in which $s(k^*)$ is known after the evaluation of Eq. (C.7). The procedure is summarized as follows.

1. Sort and renumber the m elements by increasing cost ($c_1 \leq c_2 \leq \cdots \leq c_m$).
2. Compute sequentially $L_1(k)$ [Eq. (C.7)] for $k = 1, \ldots, k^*$ until $k^* = m$ or ($L_1(k^*) < 1$ and $L_1(k^* + 1) \geq 1$). Store $s(k)$ during this loop.
3. Compute p_i by Eq. (C.4) for $i = 1, \ldots, k^*$. The value of k^* is known from step 2.
4. Set $p_i = 0$ for $i = (k^* + 1), \ldots, m$ if $k^* < m$.

[25] Recall that the elements are sorted by increasing cost value.

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

From Eq. (12), the case $r \neq 2$ and $r > 1$ is somewhat more complex,

$$\left| \begin{array}{l} \underset{\mathbf{p}}{\text{minimize}}\ \mathbf{c}^{\mathrm{T}}\mathbf{p} + \Upsilon\ (\|\mathbf{p}\|_r)^r \\ \text{subject to}\ \mathbf{e}^{\mathrm{T}}\mathbf{p} = 1 \\ \qquad\qquad\ \mathbf{p} \geq \mathbf{0} \end{array} \right. \tag{C.8}$$

where $\|.\|_r$ is the standard r-norm and we define $\Upsilon = T/(r-1)$ for convenience. This section proceeds similarly to the previous one and is based again on Miyamoto and Umayahara (1998) and Kanzawa (2013). As before, we assume that the costs c_i and the probabilities p_i are sorted by increasing value of cost.

C.2.1 The Karush–Kuhn–Tucker conditions

By proceeding as in Appendix C.1, the Karush–Kuhn–Tucker conditions are now

$$\begin{cases} \mathbf{c} + r\Upsilon\mathbf{p}^{(r-1)} - \mu\mathbf{e} - \boldsymbol{\lambda} = \mathbf{0} \\ \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{p} = 0 \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \tag{C.9}$$

where (r) is the elementwise r-power and $\boldsymbol{\lambda}$ is the Lagrange parameters vector associated with the non-negativity constraints in Eq. (C.8). As before, $\lambda_i\, p_i = 0$ for each i. A reasoning similar to the $r = 2$ case provides the following extension of Eq. (C.2)

$$p_i = \begin{cases} \sqrt[(r-1)]{\frac{1}{r\Upsilon}(\mu - c_i)} & \text{when } \mu - c_i > 0 \\ 0 & \text{when } \mu - c_i \leq 0 \end{cases} \tag{C.10}$$

However, in this case, the Lagrange parameter μ is difficult to find analytically so that the procedure derived in Appendix C.1 cannot be used. Kanzawa (2013) therefore suggested using a bisection method on the real line instead.

C.2.2 A bisection procedure

As for the $r = 2$ case, the main idea is to find numerically the optimal μ by seeking the value which exactly satisfies the sum-to-one constraint $\mathbf{e}^{\mathrm{T}}\mathbf{p} = 1$. Indeed, we observe from Eq. (C.10) that each p_i taken independently is strictly increasing with respect to the μ parameter when starting from the value $\mu = c_i$ (thus in the interval $[c_i, \infty[$) (Kanzawa 2013). Moreover, from Eq. (C.10), $p_i \geq 0$. This implies that the L1 norm of vector $\mathbf{p}$ is strictly increasing from $\mu = c_1$ (the minimum cost) in terms of μ, and

thus a bisection method (see, e.g., Press et al. 2007) 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 μ is closely approximated, Eq. (C.10) is used in order to compute the probability mass $\mathbf{p}$.

Note that Kanzawa (2013) proposes, as initial lower and upper bounds for the admissible μ,

$$\begin{cases} \mu_{\text{inf}} &= c_1 \\ \mu_{\text{sup}} &= c_{\max} + \dfrac{r\Upsilon}{m^{r-1}} \end{cases} \tag{C.11}$$

where $c_{\max} = \max_{i\in\{1,\dots,m\}}\{c_i\} = c_m$ 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 μ_{sup} necessarily leads to a L1 norm of the corresponding $\mathbf{p}$ vector greater or equal to 1. First, because $\mu_{\text{sup}} > c_{\max}$, Eqs. (C.10) and (C.11) imply that each of the m elements of $\mathbf{p}$ is strictly positive when using μ_{sup}. Then, the second expression in Eq. (C.11) can be rearranged as

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

which implies $\sqrt[(r-1)]{\frac{1}{r\Upsilon}(\mu_{\text{sup}} - c_{\max})} = 1/m$. Therefore, we must have $\sqrt[(r-1)]{\frac{1}{r\Upsilon}(\mu_{\text{sup}} - c_i)} \geq 1/m$ for each i. From Eq. (C.10), summing this expression over $i = 1, \dots, m$ shows that the L1 norm of the $\mathbf{p}$ vector associated with μ_{sup} is larger or equal to 1. The value μ_{sup} in Eq. (C.11) is therefore an upper bound for the admissible μ values. The bisection procedure is as 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 μ parameter as in Eq. (C.11).
3. Perform a bisection search on μ by computing $\|\mathbf{p}\|_1$ from Eq. (C.10) and testing if the result is lesser than or greater than 1 ($\|\mathbf{p}\|_1$ is strictly increasing in terms of μ). Stop when $\|\mathbf{p}\|_1$ is sufficiently close to 1.
4. Compute $\mathbf{p}$ from Eq. (C.10).
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 Eq. (12), extending Eq. (C.9) by taking the reference probabilities into account, $\mathbf{p}_{\text{ref}} = (p_i^{\text{ref}})$ [which correspond to $\mathbf{p}_i^{\text{ref}}$ in Eq. (12)]. The Karush–Kuhn–Tucker conditions become

$$\begin{cases} \mathbf{c} + r\Upsilon(\mathbf{p} \div \mathbf{p}_{\text{ref}})^{(r-1)} - \mu\mathbf{e} - \boldsymbol{\lambda} = \mathbf{0} \\ \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{p} = 0 \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \tag{C.12}$$

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

$$p_i = \begin{cases} p_i^{\text{ref}} \sqrt[(r-1)]{\frac{1}{r\Upsilon}(\mu - c_i)} & \text{when } \mu - c_i > 0 \\ 0 & \text{when } \mu - c_i \leq 0 \end{cases} \tag{C.13}$$

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

As in Appendix C.2, a bisection procedure can be used in order to find the probability distribution $\mathbf{p}$ summing to one. Following the derivation in Appendix C.2, the upper bound for the bisection procedure can be chosen here as $\mu_{\text{sup}} = c_{\max} + r\Upsilon$ because the reference probabilities p_i^{ref} sum to one. Thus the procedure for computing the probability distribution is exactly the same as in the previous subsection, except that Eq. (C.13) is used instead of Eq. (C.10).

References

Ahuja RK, Magnanti TL, Orlin JB (1993) Network flows: theory, algorithms, and applications. Prentice Hall, Upper Saddle River

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

Akamatsu T (1997) Decomposition of path choice entropy in general transport networks. Transp Sci 31(4):349–362

Alamgir M, von Luxburg U (2011) Phase transition in the family of p-resistances. In: Advances in neural information processing systems 24: proceedings of the NIPS 2011 conference. MIT Press, pp 379–387

Arrow K, Hurwicz L, Uzawa H (1958) Studies in linear and non-linear programming. Stanford University Press, Stanford

Barabasi AL (2016) Network science. Cambridge University Press, Cambridge

Bavaud F, Guex G (2012) Interpolating between random walks and shortest paths: a path functional approach. In: Aberer K, Flache A, Jager W, Liu L, Tang J, Guéret C (eds) Proceedings of the 4th international conference on social informatics (SocInfo'12). Lecture notes in computer science, vol 7710. Springer, pp 68–81

Beck A (2014) Introduction to nonlinear optimization. SIAM, New Delhi

Bertsekas DP (1999) Nonlinear programming, 2nd edn. Athena Scientific

Bloem M, Bambos N (2014) Infinite time horizon maximum causal entropy inverse reinforcement learning. In: Proceedings of the 53rd IEEE conference on decision and control. IEEE, pp 4911–4916

Blum M, Floyd RW, Pratt VR, Rivest RL, Tarjan RE (1973) Time bounds for selection. J Comput Syst Sci 7(4):448–461

Borg I, Groenen P (1997) Modern multidimensional scaling: theory and applications. Springer, Berlin

Brandes U, Erlebach T (eds) (2005) Network analysis: methodological foundations. Springer, Berlin

Brandes U, Fleischer D (2005) Centrality measures based on current flow. In: Proceedings of the 22nd annual symposium on theoretical aspects of computer science (STACS'05), pp 533–544

Bryson A, Ho YC (1975) Applied optimal control. Taylor and Francis, Milton Park

Buhlmann P, van de Geer S (2011) Statistics for high-dimensional data. Springer, Berlin

Busic A, Meyn S (2018) Action-constrained Markov decision processes with Kullback–Leibler cost. In: Proceedings of the 31st conference on learning theory (COLT), PMLR 75, pp 1431–1444

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

Chebotarev P (2012) The walk distances in graphs. Discrete Appl Math 160(10–11):1484–1500

Chebotarev P (2013) Studying new classes of graph metrics. In: Nielsen F, Barbaresco F (eds) Proceedings of the 1st international conference on geometric science of information (GSI'13). Lecture notes in computer science, vol 8085. Springer, pp 207–214
Chebotarev P, Shamis E (1997) The matrix-forest theorem and measuring relations in small social groups. Autom Remote Control 58(9):1505–1514
Chebotarev P, Shamis E (1998) On proximity measures for graph vertices. Autom Remote Control 59(10):1443–1459
Chung F, Lu L (2006) Complex graphs and networks. American Mathematical Society, Providence
Condat L (2016) Fast projection onto the simplex and the ℓ_1 ball. Math Program 158(1–2):575–585
Cormen T, Leiserson C, Rivest R, Stein C (2009) Introduction to algorithms, 3rd edn. MIT Press, Cambridge
Courtain S, Leleux P, Kivimki I, Guex G, Saerens M (2020) Randomized shortest paths with net flows and capacity constraints. To appear in Inf Sci
Cover T, Thomas J (2006) Elements of information theory, 2nd edn. Wiley, Hoboken
Culioli J (2012) Introduction a l'optimisation. Ellipses
Delvenne JC, Libert AS (2011) Centrality measures and thermodynamic formalism for complex networks. Phys Rev E 83(4):046117
Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
Dolan A, Aldous J (1993) Networks and algorithms: an introductory approach. Wiley, Hoboken
Doyle PG, Snell JL (1984) Random walks and electric networks. The Mathematical Association of America
Duchi J, Shalev-Shwartz S, Singer Y, Chandra T (2008) Efficient projections onto the l1-ball for learning in high dimensions. In: Proceedings of the 25th international conference on machine learning (ICML '2008), pp 272–279
Estrada E (2012) The structure of complex networks. Oxford University Press, Oxford
Estrada E, Hatano N (2008) Communicability in complex networks. Phys Rev E 77(3):036111
Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) LIBLINEAR: a library for large linear classification. J Mach Learn Res 9:1871–1874
Fouss F, Pirotte A, Renders JM, Saerens M (2007) Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. IEEE Trans Knowl Data Eng 19(3):355–369
Fouss F, Saerens M, Shimbo M (2016) Algorithms and models for network data and link analysis. Cambridge University Press, Cambridge
Fox R, Pakman A, Tishby N (2001) G-learning: taming the noise in reinforcement learning via soft updates. In: Proceedings of the 22nd conference on uncertainty in artificial intelligence (UAI 2016), pp 202–211
Françoisse K, Kivimaki I, Mantrach A, Rossi F, Saerens M (2013) A bag-of-paths framework for network data analysis. ArXiv preprint arXiv:1302.6766/v1
Françoisse K, Kivimaki I, Mantrach A, Rossi F, Saerens M (2017) A bag-of-paths framework for network data analysis. Neural Netw 90:90–111
Fred AL, Jain AK (2003) Robust data clustering. In: Proceedings of the 2003 IEEE international computer society conference on computer vision and pattern recognition (CVPR'03), vol 2, pp 128–133
Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35–41
García-Díez S, Fouss F, Shimbo M, Saerens M (2011a) A sum-over-paths extension of edit distances accounting for all sequence alignments. Pattern Recognit 44(6):1172–1182
García-Díez S, Vandenbussche E, Saerens M (2011b) A continuous-state version of discrete randomized shortest-paths. In: Proceedings of the 50th IEEE international conference on decision and control (CDC'11), pp 6570–6577
Geist M, Scherrer B, Pietquin O (2019) A theory of regularized Markov decision processes. In: Proceedings of the international conference on machine learning (ICML 2019), pp 2160–2169
Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821–7826
Griva I, Nash S, Sofer A (2008) Linear and nonlinear optimization, 2nd edn. SIAM
Guex G (2016) Interpolating between random walks and optimal transportation routes: flow with multiple sources and targets. Phys A Stat Mech Appl 450:264–277
Guex G, Bavaud F (2015) Flow-based dissimilarities: shortest path, commute time, max-flow and free energy. In: Lausen B, Krolak-Schwerdt S, Bohmer M (eds) Data science, learning by latent structures, and knowledge discovery, studies in classification, data analysis, and knowledge organization, vol 1564. Springer, pp 101–111

Guex G, Kivimaki I, Saerens M (2019) Randomized optimal transport on a graph: framework and new distance measures. Netw Sci 7(1):88–122
Guex G, Courtain S, Saerens M (2020) Covariance and correlation kernels on a graph in the generalized bag-of-paths formalism. To appear in J Complex Netw
Hashimoto T, Sun Y, Jaakkola T (2015) From random walks to distances on unweighted graphs. In: Advances in neural information processing systems 24: proceedings of the NIPS '15 conference
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin
Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity. CRC Press, Boca Raton
Havrda H, Charvat F (1967) Quantification method of classification processes. concept of structural α-entropy. Kybernetika 3(1):30–35
Hazan T, Shashua A (2007) An efficient algorithm for maximum Tsallis entropy using Fenchel-duality. Technical report TR-110, The Hebrew University of Jerusalem, Israel
Hazan T, Hardoon R, Shashua A (2007) Plsa for sparse arrays with Tsallis pseudo-additive divergence: noise robustness and algorithm. In: Proceedings of the 11th IEEE international conference on computer vision. IEEE, pp 1–8
Herbster M, Lever G (2009) Predicting the labelling of a graph via minimum p-seminorm interpolation. In: Proceedings of the 22nd conference on learning theory (COLT'09), pp 18–21
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
Ivashkin V, Chebotarev P (2016) Do logarithmic proximity measures outperform plain ones in graph clustering? In: International conference on network analysis. Springer, pp 87–105
Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106(4):620–630
Kanzawa Y (2013) Generalization of quadratic regularized and standard fuzzy c-means clustering with respect to regularization of hard c-means. In: Torra V, Narukawa Y, Navarro-Arribas G, Megías D (eds) Modeling decisions for artificial intelligence. Springer, Berlin, pp 152–165
Kanzawa Y (2018) Q-divergence-based relational fuzzy c-means clustering. J Adv Comput Intell Intell Inf 22(1):34–43
Kappen HJ, Gómez V, Opper M (2012) Optimal control as a graphical model inference problem. Mach Learn 87(2):159–182
Kapur JN (1989) Maximum-entropy models in science and engineering. Wiley, Hoboken
Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43
Keylock C (2005) Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. Oikos 109(1):203–207
Kivimäki I, Shimbo M, Saerens M (2014) Developments in the theory of randomized shortest paths with a comparison of graph node distances. Phys A Stat Mech Appl 393:600–616
Klein DJ, Randic M (1993) Resistance distance. J Math Chem 12(1):81–95
Kolaczyk ED (2009) Statistical analysis of network data: methods and models. Springer series in statistics. Springer
Kondor RI, Lafferty J (2002) Diffusion kernels on graphs and other discrete structures. In: Proceedings of the 19th international conference on machine learning (ICML'02), pp 315–322
Laha A, Chemmengath SA, Agrawal P, Khapra M, Sankaranarayanan K, Ramaswamy H (2018) On controllable sparse alternatives to softmax. In: Advances in neural information processing systems 32: proceedings of the NeurIPS'18 conference, pp 6422–6432
Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78(4):046110
Lang K (1995) Newsweeder: learning to filter netnews. In: Proceedings of the 12th international machine learning conference (ML95), pp 331–339
Lebichot B, Kivimaki I, Saerens M (2018) A bag-of-paths node criticality measure. Neurocomputing 275:224–236
Lee K, Choi S, Oh S (2018a) Maximum causal Tsallis entropy imitation learning. In: Advances in neural information processing systems 31: proceedings of the NIPS 2010 conference, pp 4403–4413
Lee K, Choi S, Oh S (2018b) Sparse Markov decision processes with causal sparse Tsallis entropy regularization for reinforcement learning. IEEE Robot Autom Lett 3(3):1466–1473
Lewis T (2009) Network science. Wiley, Hoboken
Li Y, Zhang ZL, Boley D (2011) 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). IEEE Computer Society, pp 847–856

Li Y, Zhang ZL, Boley D (2013) From shortest-path to all-path: the routing continuum theory and its applications. IEEE Trans Parallel Distrib Syst 25(7):1745–1755
Luenberger DG (1979) Introduction to dynamic systems: theory, models, and applications. Wiley, Hoboken
Luenberger DG, Ye Y (2010) Linear and nonlinear programming, 3rd edn. Springer, Berlin
Lusseau D (2003) The emergent properties of a dolphin social network. Proc R Soc Lond Ser B Biol Sci 270(2):S186–S188
Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM (2003) The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behav Ecol Sociobiol 54(4):396–405
Macskassy SA, Provost F (2007) Classification in networked data: a toolkit and a univariate case study. J Mach Learn Res 8:935–983
Manning C, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
Martins A, Astudillo R (2016) From softmax to sparsemax: a sparse model of attention and multi-label classification. In: Proceedings of the international conference on machine learning (ICML-2016), pp 1614–1623
Menard M, Courboulay V, Dardignac PA (2003) Possibilistic and probabilistic fuzzy clustering: unification within the framework of the non-extensive thermostatistics. Pattern Recognit 36(6):1325–1342
Minoux M (1986) Mathematical programming, theory and algorithms. Wiley, Hoboken
Miyamoto S, Umayahara K (1998) Fuzzy clustering by quadratic regularization. In: Proceedings of the IEEE international conference on fuzzy systems, pp 1394–1399
Miyamoto S, Ichihashi H, Honda K (2008) Algorithms for fuzzy clustering. Springer, Berlin
Muzellec B, Nock R, Patrini G, Nielsen F (2017) 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)
Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Netw 27(1):39–54
Newman MEJ (2018) Networks: an introduction, 2nd edn. Oxford University Press, Oxford
Ngyen C, Mamitsuka H (2016) New resistance distances with global information on large graphs. In: Proceedings of the 19th international conference on artificial intelligence and statistics (AISTATS'16), pp 639–647
Norris JR (1997) Markov chains. Cambridge University Press, Cambridge
Panzacchi M, Van Moorter B, Strand O, Saerens M, Kivimaki I, St Clair C, Herfindal I, Boitani L (2016) Predicting the continuum between corridors and barriers to animal movements using step selection functions and randomized shortest paths. J Anim Ecol 85(1):32–42
Peliti L (2011) Statistical mechanics in a nutshell. Princeton University Press, Princeton
Press W, Teukolsky S, Vetterling W, Flannery B (2007) Numerical recipes: the art of scientific computing, 3rd edn. Cambridge University Press, Cambridge
Price WL (1971) Graphs and networks: an introduction. London Butterworths
Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(336):846–850
Rardin R (1998) Optimization in operations research. Prentice Hall, Upper Saddle River
Reichl LE (1998) A modern course in statistical physics, 2nd edn. Wiley, Hoboken
Rubin J, Shamir O, Tishby N (2012) Trading value and information in MDPs. Springer, Berlin, pp 57–74
Saerens M, Achbany Y, Fouss F, Yen L (2009) Randomized shortest-path problems: two related models. Neural Comput 21(8):2363–2404
Schölkopf B, Smola A (2002) Learning with kernels. MIT Press, Cambridge
Silva T, Zhao L (2016) Machine learning in complex networks. Springer, Berlin
Sommer F, Fouss F, Saerens M (2016) 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, vol 9886. Springer, pp 192–201
Sommer F, Fouss F, Saerens M (2017) 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, vol 10614. Springer, pp 423–433
Strehl A, Ghosh J (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617
Tang L, Liu H (2009a) Relational learning via latent social dimensions. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'09), pp 817–826

Tang L, Liu H (2009b) Scalable learning of collective behavior based on sparse social dimensions. In: Proceedings of the ACM conference on information and knowledge management (CIKM'09), pp 1107–1116

Tang L, Liu H (2010) Toward predicting collective behavior via social dimension extraction. IEEE Intell Syst 25(4):19–25

Taylor HM, Karlin S (1998) An introduction to stochastic modeling, 3rd edn. Academic Press, Cambridge

Thelwall M (2004) Link analysis: an information science approach. Elsevier, Amsterdam

Theodorou EA, Todorov E (2012) 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). IEEE, pp 1466–1473

Theodorou EA, Krishnamurthy D, Todorov E (2013) 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

Todorov E (2007) Linearly-solvable Markov decision problems. In: Advances in neural information processing systems 19 (NIPS 2006). MIT Press, pp 1369–1375

Todorov E (2008) General duality between optimal control and estimation. In: Proceedings of 47th IEEE conference on decision and control (CDC'08), pp 4286–4292

Tsallis C (1998) Generalized entropy-based criterion for consistent testing. Phys Rev E 58(2):1442

Tsallis C (2009) Introduction to nonextensive statistical mechanics. Springer, Berlin

von Luxburg U, Radl A, Hein M (2010) 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, pp 2622–2630

von Luxburg U, Radl A, Hein M (2014) Hitting and commute times in large random neighborhood graphs. J Mach Learn Res 15:1751–1798

Wang W, Carreira-Perpinan M (2013) Projection onto the probability simplex: an efficient algorithm with a simple proof, and an application. ArXiv preprint arXiv:1309.1541 [csLG]

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Wilcoxon F (1945) Individual comparisons by ranking methods. Biom Bull 1(6):80–83

Yen L, Fouss F, Decaestecker C, Francq P, Saerens M (2007) 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). Lecture notes in artificial intelligence, vol 4426. Springer, pp 1037–1045

Yen L, Mantrach A, Shimbo M, Saerens M (2008) A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'08), pp 785–793

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

Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33(4):452–473

Zhang D, Mao R (2008a) Classifying networked entities with modularity kernels. In: Proceedings of the 17th ACM conference on information and knowledge management (CIKM 2008). ACM, pp 113–122

Zhang D, Mao R (2008b) A new kernel for classification of networked entities. In: Proceedings of 6th international workshop on mining and learning with graphs, Helsinki, Finland