Perri et al. *EPJ Data Science* (2023) 12:48
https://doi.org/10.1140/epjds/s13688-023-00416-3

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Bayesian inference of transition matrices from incomplete graph data with a topological prior

Vincenzo Perri[1†], Luka V. Petrović[1*†] and Ingo Scholtes[2,1]

*Correspondence:
petrovic@ifi.uzh.ch
[1]Data Analytics Group, Department of Informatics, University of Zurich, Binzmühlestrasse 14, CH-8050 Zurich, Switzerland
Full list of author information is available at the end of the article
[†]Equal contributors

Abstract

Many network analysis and graph learning techniques are based on discrete- or continuous-time models of random walks. To apply these methods, it is necessary to infer transition matrices that formalize the underlying stochastic process in an observed graph. For weighted graphs, where weighted edges capture observations of repeated interactions between nodes, it is common to estimate the entries of such transition matrices based on the (relative) weights of edges. However in real-world settings we are often confronted with incomplete data, which turns the construction of the transition matrix based on a weighted graph into an *inference problem*. Moreover, we often have access to additional information, which capture topological constraints of the system, i.e. which edges in a weighted graph are (theoretically) possible and which are not. Examples include transportation networks, where we may have access to a small sample of passenger trajectories as well as the physical topology of connections, or a limited set of observed social interactions with additional information on the underlying social structure. Combining these two different sources of information to reliably infer transition matrices from incomplete data on repeated interactions is an important open challenge, with severe implications for the reliability of downstream network analysis tasks.

Addressing this issue, we show that including knowledge on such topological constraints can considerably improve the inference of transition matrices, especially in situations where we only have a small number of observed interactions. To this end, we derive an analytically tractable Bayesian method that uses repeated interactions and a topological prior to perform data-efficient inference of transition matrices. We compare our approach against commonly used frequentist and Bayesian approaches both in synthetic data and in five real-world datasets, and we find that our method recovers the transition probabilities with higher accuracy. Furthermore, we demonstrate that the method is robust even in cases when the knowledge of the topological constraint is partial. Lastly, we show that this higher accuracy improves the results for downstream network analysis tasks like cluster detection and node ranking, which highlights the practical relevance of our method for interdisciplinary data-driven analyses of networked systems.

Keywords: Pairwise interactions; Weighted graphs; Network inference; Bayesian learning; Data efficiency

Springer

1 Introduction

Graph models of relational data have become a cornerstone in the analysis of complex systems (Boccaletti et al. [2]) and an important foundation for the application of machine learning to graph-structured data from social, technical, and biological systems (Bronstein et al. [3]). Many network analysis and graph learning techniques are based on discrete- or continuous-time models of random walks (Masuda et al. [21]) such as, e.g., community detection algorithms like InfoMap (Rosvall and Bergstrom [35]) or WalkTrap (Pons and Latapy [32]), node ranking techniques like PageRank (Page et al. [24]), neural graph embeddings like DeepWalk (Perozzi et al. [30]) or node2vec (Grover and Leskovec [9]), walk-based similarity scores that are the basis for link prediction (Liben-Nowell and Kleinberg [18]), or heat kernels for graphs used for community detection (Kloster and Gleich [14]) and node ranking (Chung [7]). To apply these methods, it is necessary to obtain a transition matrix that formalizes the underlying stochastic process in the observed graph. This is trivial when we have full information on repeated interactions in the graph, which enables us to estimate transition probabilities between nodes based on relative frequencies of observed interactions. However in real-world settings we are often confronted with incomplete data, which turns the construction of the transition matrix into an *inference problem* that we need to address to obtain reliable results.

In this work, we consider situations where we have access to a possibly incomplete set of observed repeated interactions between a set of nodes. Such data can be represented as a weighted graph, where the weight of an edge corresponds to the number of observed interactions between a given node pair. For such weighted graphs, it is common to define transition probabilities proportional to the edge weights. From a statistical inference point of view, this method to infer the transition matrix based on observed interactions corresponds to a frequentist approach that uses a maximum likelihood estimation. When few observations are available, this simple approach suffers from overfitting: On the one hand, unobserved interactions translate to zero transition probabilities even though transitions may actually be possible in the underlying graph. On the other hand, those interactions that were observed are likely to translate to overestimated transition probabilities. This generates a large variance in inferred transition probabilities that can severely distort the results of downstream network analysis and graph learning tasks.

However, repeated interactions are often not the only information that we have about the networked systems that we study. We often have additional knowledge about topological constraints that determine which of the interactions are theoretically possible and which others are not. For example, consider a transportation system, where observed interactions represent passengers travelling between connected stations, or a social network, where interactions represent messages transferred between users, or a Web graph, where interactions represent users clicking on a hyperlink between two Web pages. In the first case, the movement of passengers is constrained by the available transportation infrastructure, in the second, the spreading of information is constrained by existing social connections, while in the third case, available hyperlinks constrain the possible clickstreams of users. Such constraints can limit the number of parameters of the model that we seek to infer. They can thus help us to address the reliable inference of transition matrices and improve the results of downstream analyses (see steps for the analysis of network data illustrated in Fig. 1). Therefore, in this paper, we address the following research questions:

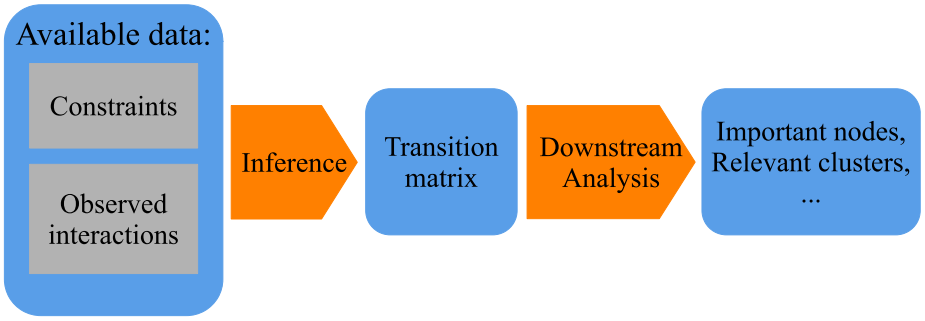

Figure 1 Steps in the analysis of network data. The first step is the inference of the transition matrix from the network data. In the second step, the transition matrix is used to detect the important nodes, relevant clusters and other interesting features of the system. The goals of our work are to extend the network inference to use the topological constraints, to investigate the robustness of the new method to partial knowledge of the constraints, and to investigate whether this improves downstream analysis tasks

Q1 How can we use the information on topological constraints to improve the inference of transition matrices from incomplete data on repeated interactions captured in weighted graphs?

Q2 How is the inference of transition matrices influenced by a *partial* knowledge of the underlying topological constraints, which is often the case in real-world settings?

Q3 To what extent can our proposed approach of including topological constraints in the inference of transition matrices improve the performance of downstream network analysis tasks like node ranking and community detection?

The remainder of this article is structured as follows: In Sect. 2, we formally define the inference problem that we address in our work and we introduce two methods that are commonly used to infer transition matrices without leveraging topological constraints, namely maximum likelihood estimation and a noninformative Bayesian approach. Addressing Q1, in Sect. 3 we apply Bayesian learning to the problem of inferring transition matrices in incomplete data on repeated interactions in a graph. We include the topological graph constraints in the prior, and call this method BaCon. Furthermore, we explain how to select the "shape" hyperparameter. In Sect. 4, we introduce the datasets and the experimental setup that we use to evaluate our method. We next compare BaCon to the methods introduced in Sect. 2 (i.e. frequentist and noninformative Bayes approach) that do not use information on topological constraints. In Sect. 5, we evaluate the extent to which the inclusion of the topological constraints improves the network inference, and address Q2, observing the effects of partial knowledge of the constraint. In Sect. 6, we address Q3 and explore whether the effects of the network inference carry over to the network analysis results. In Sect. 7, we show how inference of diverse examples of real-world networks can be further improved with an appropriate choice of the shape hyperparameter, and demonstrate the effects of model selection of the shape hyperparameter. In Sect. 8 we discuss how our work complements existing network analytic methods in the field. We finally summarize our conclusions and outline future work in Sect. 9.

The results of our study show that (i) the inclusion of topological constraints considerably improves the inference of transition matrices in network data, and (ii) that this improved inference translates to increased accuracy for downstream network analysis tasks. Our work highlights the importance of treating the construction of network mod-

els from (partially observed) interactions as an inference problem that can be addressed using Bayesian statistics. This is in line with the work of Peel et al. [25], who note that the standard inference methods in network science have little consideration for inherent uncertainties and incompleteness of readily-available empirical data, or even an in-depth understanding of its provenance and measurement procedure. Our results further suggest that constraints for the topology of interactions in complex systems should be given more attention as their inclusion can significantly improve modeling accuracy, especially in the limit of small data. Given the importance of network inference in partially observed or noisy data we expect our work to be of interest for a broad interdisciplinary community. An implementation of our method is available as an Open Source project (Zenodo [44]).

2 Background

This section introduces the mathematical formalism for the inference of transition matrices from observations on repeated interactions and outlines two existing strategies for addressing it. We assume that we observe a multiset of pairwise interactions (i,j) between nodes $i,j \in V$, we denote it with $\mathcal{E}$ and indicate the number of times the interaction (i,j) occurred with n_{ij}. In addition, we are given a directed graph $G = (V,E)$ that represents the topological constraints, i.e., which edges can be observed and which are impossible. Given these sources of data, our goal is to infer the transition matrix $\mathbf{T}$ that best reflects the actual transition probabilities in the system.

The simplest option is to assign the transition probabilities in a frequentist way, i.e. proportionally to the observation counts n_{ij} and thus $\pi_{ij} = n_{ij}/\sum_{k\in V} n_{ik}$. This corresponds to a maximum likelihood estimation (MLE) of a multinomial distribution $p(j|i)$. The advantages of this approach are its simplicity and good performance when we have a data set that is sufficiently large considering the space of possible interactions. In fact, this method is so common that it has become a standard "preprocessing" step, which is rarely even mentioned as a method to "infer" a graph model from observational data. However, this simple method can require a large number of observations to assign non-zero probabilities to all edges that are possible based on the underlying topology of the system. In a nutshell, when few observations are available, it is difficult to determine whether an edge with probability zero is not possible or whether it has not been observed yet. From a machine learning point of view, the erroneous interpretation of edges with zero probability as evidence of the absence of the associated edge corresponds to an *overfitting* of the weighted graph model.

An alternative to a frequentist approach is to use a Bayesian one. The noninformative Bayesian approach addresses the issue of overfitting by recording the distribution of parameters π_{ij} for a given data set. For every node i, we organize the parameters π_{ij} in vectors $\vec{\pi}_i = (\pi_{ij})_{j\in V}$. A priori, we would assign, e.g., a uniform prior over the space of transition probabilities: $p((\vec{\pi}_i)_{i\in V}) = const. = \prod_{i\in V} \mathrm{Dir}(\pi_i|\vec{\alpha}_i = \vec{1}_N)$, where Dir denotes the Dirichlet distribution, $\vec{\alpha}_i$'s are its concentration parameters, and $\vec{1}_N$ is the N-dimensional vector with all components equal to 1. This choice of parameters corresponds to the uniform distribution. The noninformative Bayesian approach further uses Bayes' rule to update the prior distribution of transition probabilities: $p((\vec{\pi}_i)_{i\in V}|\mathcal{E}) = p(\mathcal{E}|(\vec{\pi}_i)_{i\in V})p((\vec{\pi}_i)_{i\in V})/p(\mathcal{E})$. An advantage of this method is that unobserved edges are still modeled with non-zero probabilities. However, this also introduces problems: First, since all transitions are modelled with non-zero probabilities, we cannot directly use sparse matrices, which complicates

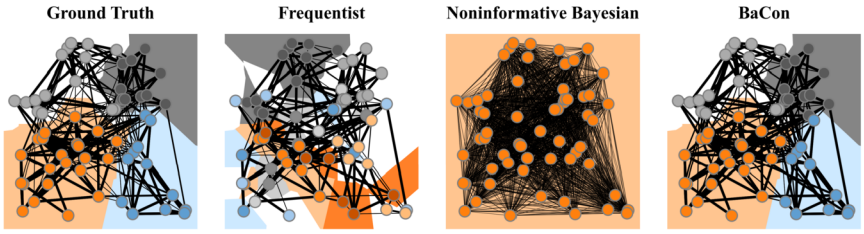

Figure 2 Illustration of methods to infer transition matrices from repeated interactions on a downstream task of clustering. The first panel shows the ground truth clusters encoded in the transition probabilities of a geometric random network with Euclidean metric in which we observe small set of interactions. We use three different inference methods to construct a transition matrix and detect communities using InfoMap. The other panels show the clusters detected using the transition matrix inferred with a frequentist, noninformative Bayesian, and our approach - BaCon. BaCon finds clusters that are closer to the ones detected from the ground truth matrix

applications to large networks. Second, and more importantly, in the typical case of networks with sparse topologies, a large amount of data is required to overcome the uniform prior on a fully connected graph. From a machine learning point of view, this corresponds to an *underfitting* of the weighted graph model.

In the first three panels of Fig. 2, we use a toy example to illustrate the issues of frequentist and noninformative Bayesian network inferences. In the toy example, we consider a ground-truth transition matrix constrained to a random geometric graph topology (first panel). The network consists of four clusters expressed in higher transition probabilities between nodes within the same cluster. We draw a small random sample of repeated interactions from the distribution of edge probabilities in the ground truth network. We apply the frequentist (second panel) and the noninformative Bayesian method (third panel) to infer a transition matrix. We further visualize the result of the popular community detection technique InfoMap on the resulting transition matrix. The frequentist approach detects many spurious communities because of its overfitting issue. The Bayesian approach detects a single community because the observed data is insufficient for overcoming the prior.

Neither the frequentist approach, nor the noninformative Bayesian, use the information on the constraints $G = (V, E)$. Addressing this issue, in the next section, we formally introduce a Bayesian method that leverages topological constraints (BaCon) that are often known in real-world networked systems. In the fourth panel of Fig. 2, we show a representative example that illustrates how inclusion of such constraints in the network inference improves the detection of clusters.

3 BaCon: Bayesian constrained network inference

In this section, we show how the knowledge of topological constraints can be used as a prior to infer transition probabilities from observations of interactions (Q1). We formally introduce a Bayesian method (BaCon) that leverages topological constraints that are often available for real-world networked systems.

We assume that we are given a set of nodes V of size N, and a set of possible edges $E \subset V \times V$ that represent a topological constraint for our inference task. I.e. we assume that an interaction between node i and node j cannot occur if $(i,j) \notin E$. For each node i, we denote its possible successors as $S(i)$. We also assume that we are given a data set

$\mathcal{E}$ of interactions that were observed in the system. Our goal is to infer the best directed weighted graph model from this data set. The transition probabilities $p(j|i) = \pi_{ij}$ from i to its successors $j \in S(i)$ satisfy:

$$\sum_{j \in S(i)} \pi_{ij} = 1, \quad \forall i, j : 0 < \pi_{ij} < 1. \tag{1}$$

We are interested in the probability density of parameters π_{ij}. We again organize them in vectors $\vec{\pi}_i$, similarly to the noninformative Bayesian approach introduced in the previous section, but this time only over the possible successors: $\vec{\pi}_i = (\pi_{ij})_{j \in S(i)}$. We capture the information about impossible transitions using a Dirac delta function as prior for the π_{ij}. Since those transitions can not occur in the data, their probability distribution will not change. For the possible transitions from a node i, we assume a uniform prior over the parameter space:

$$p(\vec{\pi}_i|G) = \mathrm{Dir}(\vec{\pi}_i|\vec{\alpha}_i = \alpha \vec{1}_{|S(i)|}), \tag{2}$$

where, for the uniform distribution, we choose $\alpha = 1$.

The choice of $\alpha > 1$ means that the equiprobable distributions are more likely, and the choice of $\alpha < 1$ means that the biased distributions are more likely. We will explore the choice of α in Sect. 7. In the following we denote all transition probabilities as $\pi := (\vec{\pi}_i)_{i \in V}$. The likelihood of observing data $\mathcal{E}$, which contains n_{ij} observations of an edge (i, j), is given by the multinomial distribution:

$$p(\mathcal{E}|\pi, G) = Z \prod_i \prod_{j \in S(i)} \pi_{ij}^{n_{ij}}, \tag{3}$$

where Z denotes the number of permutations of observations. We use Bayes' rule to update our prior distribution into a posterior after observing data $\mathcal{E}$:

$$p(\pi|\mathcal{E}, \alpha, G) = \frac{p(\mathcal{E}|\pi, G) \times p(\pi|\alpha, G)}{p(\mathcal{E}|\alpha, G)}. \tag{4}$$

As the multinomial and the Dirichlet distributions are conjugate distributions, the posterior is also a Dirichlet distribution with parameters $\vec{\alpha}_i = \vec{\alpha}_i^0 + \vec{n}_i$, where $\vec{\alpha}_i^0$ denotes the a priori concentration parameters, and $\vec{n}_i := (n_{ij})_{j \in S(i)}$. The posterior $p(\vec{\pi}_i|\mathcal{E}) = \mathrm{Dir}(\vec{\pi}_i|\vec{\alpha}_i = \alpha \times \vec{1}_{|S(i)|} + \vec{n}_i)$ defines an ensemble of parameters consistent with the observations. From this ensemble we can compute the expected value of the transition matrix as:

$$\mathbf{E}[T_{ij}] = \frac{\alpha_{ij}}{\sum_k \alpha_{ik}}.$$

We will use the expected value of the transition matrix in our experiments where we compare the outlined inference method versus the alternatives.

We showed a method to infer the transition matrix $\mathbf{T}$ from observed interactions between system elements, using the topological information as a key ingredient of the prior distribution, which provides an answer to the first research question. We highlight that the method has one free parameter α, which governs whether the prior distribution of the

transition probabilities favours equiprobable or biased distributions. In our experiments, we chose $\alpha = 1$ which corresponds to the uniform prior. However, in Sect. 7, we show the negative impact that this choice can have on the inference. In the following, we describe how to select the parameter α from a several candidates, and we experimentally test the model selection on the two real world datasets for which this approach is of interest.

To select the parameter α, we assume that we do not have a preference between the candidates $\{\alpha_1, \alpha_2, \ldots, \alpha_l\}$, i.e. we start with a uniform distribution. Thus, the likelihood of the hyperparameter α_i, $p(\mathcal{E}|\alpha_i, G)$, which is exactly equivalent to the marginal likelihood in Eq. (4) and can be computed analytically, is directly proportional to the posterior probability of the hyperparameter α_1. Finally, we select the maximum a posteriori α_i. For further reading of this approach, see Kass and Raftery [11].

4 Datasets

In this section, we describe the synthetic and empirical datasets that we use to test BaCon and investigate the research questions Q2 and Q3. For the synthetic data-sets, we generate the ground truth transition matrix using three different random graph models to define the underlying constraints. To generate the transition probabilities, we use a simple procedure for the transition matrix and centrality ranking comparison and a slightly more complex one for the task of community detection. We further use five real-world datasets from ecology, neurology, information systems, transportation systems and technical systems.

4.1 Synthetic datasets

In this section we describe the procedure that we use to generate the synthetic data. First, we generate the underlying topology $G = (V, E)$, then, based on this topology, we randomly generate transition probabilities of a ground truth transition matrix $\mathbf{T}_{\text{gt}}$, and finally, we use the transition matrix to generate the synthetic interactions $\mathcal{E}$. This synthetic generation process is necessary to ensure that the target pattern (e.g. the ground truth clustering or the ground truth ranking) cannot be fully recovered only from the graph topology $G = (V, E)$, and thus that the ground truth probabilities are necessary for the downstream task.

Underlying topology To generate the underlying topology for our synthetic experiments, we first generate an undirected network using one of three different generative models: (1) an Erdős-Rényi $G(N, M)$ random graph model, (2) soft random geometric graphs with Euclidean metrics of latent spaces, and (3) soft random geometric graphs with hyperbolic metrics of latent spaces. The Erdős-Rényi $G(N, M)$ model (where $N = 500$ is the number of nodes and $M = 5000$ the number of edges) is one of the simplest random graph models. It generates networks with the small distance between pairs of nodes that characterizes real networks, but without their degree heterogeneity, high clustering values, and modular structures. In contrast, soft random geometric graphs have high clustering and display the emergence of modular structures. They generate networks with different degree distributions depending on the metric of the latent space. In our experiments, we use both a Euclidean metric, which generates networks with high clustering but uniform degree distributions, and a hyperbolic metric, which generates networks with high clustering and power-law degree distributions (Krioukov et al. [16]). Each of the generative models isolates distinct aspects of the topology of real-world networks and allows us to test the effect of these aspects on the inference problem. To produce a geometric graph,

we scatter $N = 500$ nodes in a two-dimensional latent space. For the Euclidean metric, we uniformly scatter nodes in a $(0,1) \times (0,1)$ square. For the hyperbolic metric, we uniformly scatter nodes in a hyperbolic disk of radius $R = 1$ and constant Gaussian curvature $-\zeta^2$ with $\zeta = 1$. We connect nodes i and j with probability $p((i,j)) = \exp(-d_{ij}/\sigma)$, where d_{ij} is the pairwise distance of nodes in the latent space and $\sigma = 0.1$ is the scale parameter. Large values of σ indicate a higher probability of connecting to distant nodes, leading to denser networks. Low values of σ favor connections with closer nodes and lead to sparser topologies. Having generated the network, we add an edge from a random node of every connected component to a random node in the largest connected component, thus enforcing the connectedness of the network. Finally, we represent the undirected network as a directed network by converting each undirected edge (i,j) to two directed edges (i,j) and (j,i). For this choice of parameters, Erdős-Rényi networks contain around 10,000 directed edges, Euclidean networks produce around 12,000 directed edges and hyperbolic networks produce around 4000 directed edges. Obtaining the network constraint from a connected undirected network ensures that random realizations of the topology have no sink nodes, which simplifies the experiments by avoiding potential pathological cases.

Ground truth transition matrix Next, we set the probabilities on the underlying topology, thus defining the ground truth model. To generate the ground truth transition matrix $\mathbf{T}_{\text{gt}}$, we generate the transition probabilities π_{ij} of outgoing edges for each node i from $(\pi_{ij})_{j \in S(i)} \sim \text{Dir}(\vec{\alpha}_i)$, with two different choices of α_i.

For the experiments on transition matrix inference and on node ranking, we choose a uniform distribution: $\vec{\alpha}_i = \vec{1}_{S(i)}$. However, to ensure that we do not produce networks with a single community, we employed the following procedure. We first artificially group the nodes, and choose a larger value for the component α_{ij} of $\vec{\alpha}_i$ when j is in the same community as i ($\alpha_{ij} = 10$) then when it is not ($\alpha_{ij} = 1$). We group the nodes of Erdős-Rényi graph by uniformly placing nodes in three groups; in the case of the random Euclidean graph, we cut the latent space with a random horizontal and a random vertical line, and assign nodes to the same group if they belong to the same partition of the latent space; in the case of random hyperbolic graph, the procedure is the same, except that we cut the space with two geodesics. By grouping nodes in this way, we ensure that the topology of the geometric graphs holds some information about the cluster structures, but that the probabilities are not irrelevant for detecting the cluster structure.

Edge sampling In the third step, we generate a sample of interactions $\mathcal{E}$. Since we do not expect the starting nodes to be observed with uniform frequency, we assign random "starting probabilities" to different nodes. We draw the starting probabilities from $\text{Dir}(\vec{\alpha} = \vec{1}_N)$. We choose the starting node i as a multinomial draw of the starting node probability distribution. We choose the successor j as a multinomial draw of the probability distribution of successors. The sampling simulates the observation of interactions in a networked system.

4.2 Empirical datasets

In this section we describe the empirical weighted networks that we used to evaluate the methods. The five real-world empirical networks come from different domains.

The first network captures an *ecosystem*, which is the food chain of Florida bay (Verasztó et al. [40]). It contains 2106 edges capturing carbon exchange between 128 species. An

entry A_{ij} of the adjacency matrix represent the biomass exchanged from species i to species j (i.e. j ate A_{ij} quantity of i). We assume that the non-zero values of the adjacency matrix define the possible edges.

The second network, denoted as *neural*, in the domain of neuroscience, captures the connectome (i.e. the map of neural connections) of a larva of a simple marine worm (Ulanowicz and DeAngelis [39]). The topology is defined with 11,437 axons connecting 2728 neurons. The entry A_{ij} of the adjacency matrix is the number of synapsis connecting neuron i to neuron j.

The third network is the Wikipedia web graph used to play the *wikispeedia* game in West and Leskovec [42]. In the game, players have to find a short path between two pages in the graph. The topological constraint is defined by the 239,764 hyperlinks that the players could use to navigate through 4592 Wikipedia pages. An entry of the adjacency matrix A_{ij} is the number of times users clicked on the hyperlink from i to j while playing the game.

The fourth network records *flights* between US airports. The data (TransStat [38]) captures 286,810 passenger itineraries between 175 US airports as recorded in 2014. An entry of the adjacency matrix A_{ij} is the number of passengers that flew from i to j, and we assume that the connection does not exist if none of the observed passengers used it.

The fifth network is the central Chilean *power-grid* (Kim et al. [13]). The topological constraint is defined from 444 connections between 347 stations. The entry of the adjacency matrix A_{ij} is the capacity of the connection measured in Kilovolts. Although the *powergrid*, *neural*, and *ecosystem* data sets do not quite fit in our setting (e.g. connection voltage does not determine how often we observe the connection), we can still use them to test the inference methods on a diverse set of real-world networks.

Each of the empirical networks defines the adjacency matrix $\mathbf{A}$. We define the underlying constraint $G = (V, E)$ as a graph of non-zero weighted edges $(i, j) \in E \Leftrightarrow A_{ij} > 0$. The ground truth transition matrix $\mathbf{T}_{\text{gt}}$ is obtained by normalizing rows of the adjacency matrix $\mathbf{A}$. We simulate observations by sampling edges from the adjacency matrix $\mathbf{A}$ of the corresponding weighted network. We put all edges (i, j) in a single bin and draw them with probability proportional to A_{ij}. The sampling simulates an input of interactions observed on a networked system. Since the sampling introduces variability on the observed edges, we run the experiments multiple times.

5 Effects on inference of weighted graphs

We now explore the effect of using topological constraints in the inference of transition matrices. We compare the proposed method BaCon, which uses information on the topological constraint, to the frequentist and the noninformative Bayesian approaches introduced in Sect. 2, which do not incorporate topological information. At the end of the section, we investigate the research question Q2 and explore how partial knowledge of topological constraints influences the inference.

Experimental setup The input data consists of a sample of observed edges $\mathcal{E}$ and a given graph constraint G as described in Sect. 4. The target variable is the ground truth transition matrix $\mathbf{T}_{\text{gt}}$, which we construct as explained in Sect. 4. We use the data to infer a transition matrix $\mathbf{T}$ using each method and compare those transition matrices to the ground truth $\mathbf{T}_{\text{gt}}$. We quantify the errors in estimating the matrix entries by computing the Frobenius norm of the difference between the inferred transition matrix and the ground

truth transition matrix $\|\mathbf{T}_{\text{gt}} - \mathbf{T}\|$. Since the Frobenius norm of a matrix $\mathbf{X}$ is defined as $\|\mathbf{X}\| = (\sum_{ij} X_{ij}^2)^{1/2}$, computing the Frobenius norm of the difference between transition matrices is equivalent to computing the mean square error of the transition probabilities inferred by the models. The worst value for $\|\mathbf{T}_{\text{gt}} - \mathbf{T}\|$ is $2N$ where N is the number of nodes in G; the best is 0, which indicates a perfect match between the inferred matrix and the ground truth transition matrix. We evaluate the performance of methods for different sizes of the sample $\mathcal{E}$ and we plot how the error of the inferred transition matrix depends on the number of sampled edges $\mathcal{E}$. We run the experiment 100 times for each sample size. In all figures, we represent the average error of the frequentist method with the dotted black line, the dashed dark-blue curve shows the average error of the noninformative Bayesian method. The average error of BaCon is represented with the solid orange curve. As a reference point, we drew the light-gray dashed horizontal line, which represents the average difference between the ground truth transition matrix and the transition matrix constructed based on the unweighted topology, which does not depend on the sampled edges. We represent the variability in the results' distribution using vertical error bars that indicate the 95% confidence intervals (specifically, the intervals between the 2.5-th and the 97.5-th percentiles).

Results We present our results on the dependency between the inference error (y-axis) and the sample size of interactions (x-axis) in Fig. 3 for synthetic networks and in Fig. 4 for empirical networks. The results show that the inclusion of the topological constraint positively affects the inference of the transition matrix. All three methods show an improvement in performance with the number of sampled edges and converge to the ground truth transition matrix for a large number of observations. However, their behavior differs greatly in the intermediate and small data range. When few edges are available, the frequentist and noninformative Bayes perform similarly, and both do worse than the unweighted topology. In the intermediate range, the two methods display very different behaviors. The frequentist approach exhibits a peak in its error (particularly marked for the synthetic experiments), indicating that the method over-adjusts to the observations. On the contrary, the noninformative Bayesian method steadily improves performance but at

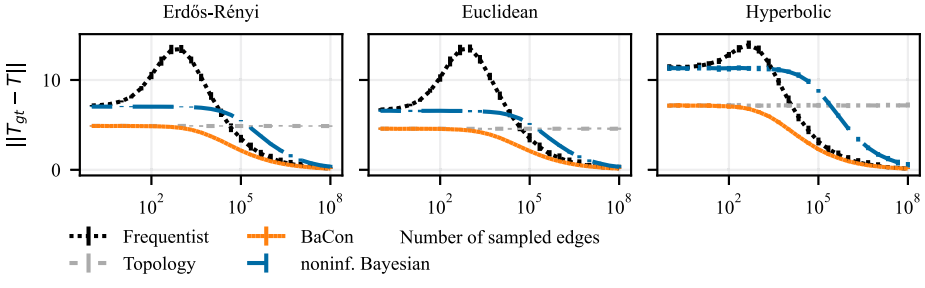

Figure 3 Effects of the use of constraints on the inference of a transition matrix in synthetic data. The left panel shows results on Erdős Rényi graphs, the middle panel shows results on geometric Euclidean random graphs, and the right panel shows results on geometric hyperbolic random graphs. We measure performance with the Frobenius distance between the transition matrix of the ground truth and that of the inferred model. Error bars represent the 95% confidence interval (mostly too small to be visible). The inclusion of the constraints allows BaCon to recover the transition matrix more precisely than both the frequentist and the noninformative Bayesian approaches

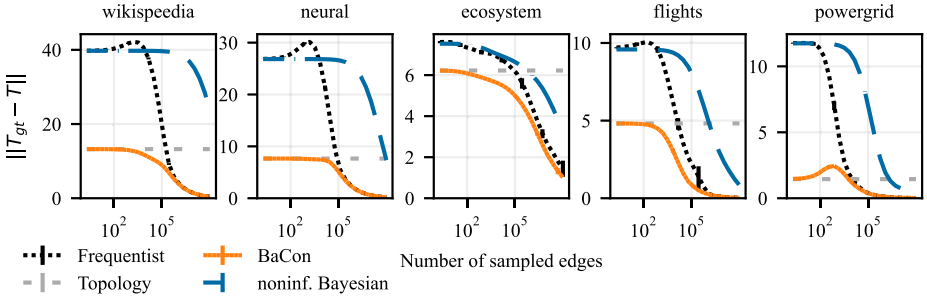

Figure 4 Effects of the use of constraints on the inference of a transition matrix in empirical data. From left to right, the panels show the results for *wikispeedia*, *neural*, *ecosystem*, *flights*, and *powergrid* networks. We measure performance with the Frobenius distance between the transition matrix of the ground truth and that of the inferred model. Error bars represent the 95% confidence interval (mostly too small to be visible). In accordance with the synthetic results, the inclusion of the constraints allows BaCon to recover the transition matrix more precisely than both the frequentist and the noninformative Bayesian approaches

a very slow pace: due to the use of a fully connected prior (i.e. the lack of topological information) each observation only provides very little information for the inference. As a result, the frequentist method surpasses the noninformative Bayesian when the number of observed edges increases despite its peak in error in the intermediate data range. BaCon mediates between these two behaviors. Its inference begins from a better initial performance by relying on the constraint provided by the unweighted topology. In the intermediate data range, BaCon's performance improves considerably faster than the noninformative Bayesian approach because, due to the topological constraints, it has less degrees of freedom. The constraint also gives BaCon the necessary information to avoid overreliance on the observations, thus preventing the peak in error that characterizes the frequentist method. BaCon steadily improves its performance when more data becomes available, converging to the ground truth transition matrix faster than the other approaches.

To give an idea of how large these improvements can be, an aspect which might be hidden by the logarithmic axes, we highlight the difference in the amount of data required by the different methods to reach a fixed mean square error of 2.5 in synthetic data. The frequentist method needed over three times more data than BaCon in the case of Erdős-Rényi random graphs, over four times more data in Euclidean random graphs and double the amount of data in the case of hyperbolic random graphs. For the noninformative Bayesian method the ratios are even more pronounced: it needed over twenty times more data than BaCon in the case of Erdős-Rényi random graphs and Euclidean random graphs and over fifty times more data in the case of hyperbolic random graphs.

Discussion With these experiments, we have investigated the impact of including topological constraints on the inference of a transition matrix. The results highlight that the inclusion of constraints provides a major improvement in data-efficiency. We noticed both the overfitting of the frequentist and the underfitting of the noninformative Bayesian approaches introduced in Sect. 2. The inclusion of constraints provides an alternative source of information in the limit of small data sizes which prevents overfitting. It also reduces the degrees of freedom, which prevents underfitting.

It is worth mentioning that the Frobenius norm, which we used to evaluate the inferred matrix, is not the only matrix norm we tried. However, the other matrix norms showed

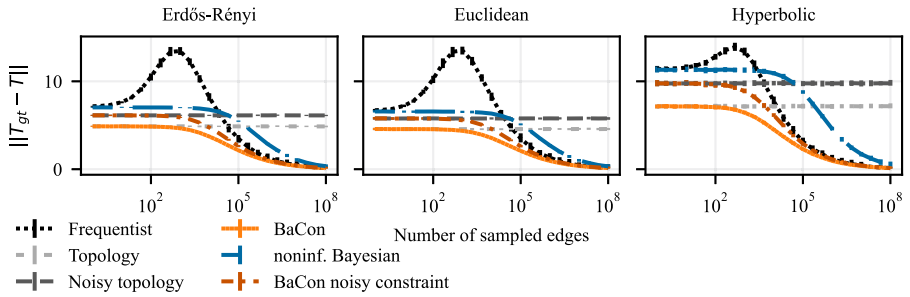

Figure 5 Effects of partial constraints on the inference of a transition matrix. The left panel shows results on Erdős Rényi graphs, the middle panel shows results on geometric Euclidean random graphs, and the right panel shows results on geometric hyperbolic random graphs. We measure performance with the Frobenius distance between the transition matrix of the ground truth and that of the inferred model. Error bars represent the 95% confidence interval (mostly too small to be visible). Despite the partial information on constraints, its usage benefits the network inference

the same trends, and we selected the Frobenius norm because of the intuitiveness of its connection to the mean square error. The results for the other matrix norms are available in the complementary Zenodo package (Zenodo [44])

One potential threat to the validity of these results is that the constraint we used in the inference is identical to the constraint we used to generate the data. However, assuming full knowledge of the constraints is unrealistic for real-world systems. In the next section, we tackle this case to check whether missing information about constraints invalidates the BaCon approach.

Sensitivity to partial constraints We now investigate research question Q2 and test the robustness of our inference method against incomplete knowledge about the underlying topology. We introduce noise to the constraints by adding spurious edges alongside those that capture the underlying topology. We note that the addition of such spurious edges effectively corresponds to a loss of information on the constraint. An addition of all possible edges corresponds to the noninformative Bayesian method with a fully connected prior introduced in Sect. 2. Aiming for an equal number of original and spurious edges, we add 10,000 uniformly random spurious directed edges to the Erdős-Rényi random graphs, 12,000 to geometric Euclidean random graphs, and 4000 to the geometric hyperbolic random graphs.

The results in Fig. 5 display that missing information on the constraints decreases the performance of BaCon in the small data regime, and that it needs more data to make up for that deficiency. However, the method still outperforms the other two approaches, which do not use information about constraints at all. Answering Q2, this demonstrates that the proposed method BaCon is robust to missing information about topological constraints in complex systems.

6 Effects on downstream analyses of weighted graphs

In this section, we investigate research question Q3 and measure the effects of the inference on the downstream tasks of node ranking and clustering.

Experimental setup To evaluate the impact of the inference methods on downstream network analysis tasks, we perform experiments that build on the experimental setup from

Sect. 5. The input data again consists of the sample of edges $\mathcal{E}$ and a graph constraint G, as described in section Sect. 4. From the input data, we infer the transition matrix using all three inference methods. However, instead of evaluating the error of the inferred matrix $\mathbf{T}$, in these experiments we perform downstream tasks of node ranking and cluster detection using the inferred transition matrix. We use PageRank to compute a node ranking on $\mathbf{T}$, and compare it to the corresponding PageRank node ranking obtained based on the ground truth transition matrix $\mathbf{T}_{\text{gt}}$. We measure the performance using the Kendall-τ correlation between the two rankings; a perfect match of the rankings leads to $\tau = 1$, uncorrelated rankings lead to $\tau \approx 0$, whereas the reverse ranking is captured by $\tau = -1$. To detect clusters, we apply the random-walk based community detection algorithm Infomap (Rosvall and Bergstrom [35]). This method is a convenient choice for our analysis as it naturally detects clusters in directed networks along with the optimal number of clusters. We measure the performance by computing the adjusted mutual information (AMI) between the clusters detected in the inferred transition matrix $\mathbf{T}$ and the clusters detected using the ground truth transition matrix $\mathbf{T}_{\text{gt}}$. A perfect agreement has an AMI score equal to 1, completely unrelated assignments have AMI of 0. We run 100 independent experiments for each sample size and each task. We plot the average performance of downstream tasks as a function of the size of the edge sample $\mathcal{E}$. To represent the variability of the performance, we show the 95% confidence intervals as the error bars. The average performance using the frequentist inference is presented with the dotted black curve, the dashed blue is for the noninformative Bayesian method, and the solid orange is showing the performance of BaCon. The dashed light-gray horizontal line, which we show as a reference point, again represents the average performance obtained using only the unweighted topology.

Node ranking We present the results of the node ranking experiments on synthetic networks in Fig. 6, and on the empirical networks in Fig. 7. Except for the *ecosystem* dataset, the inclusion of constraints in network inference improves the ranking of nodes. Both the frequentist and the noninformative Bayesian methods start from a performance of $\tau = 0$ in the limit of very small data sizes, and need considerable data to reach the performance of the unweighted topology. On the other hand, BaCon defaults to the performance of the topology in the limit of small data. BaCon performs at least as good as the unweighted

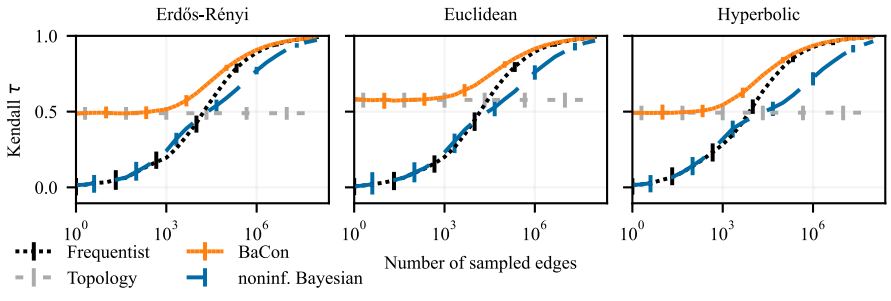

Figure 6 Effects of constraints on the downstream task of node ranking. The left panel shows results on Erdős Rényi graphs, the middle panel shows results on geometric Euclidean random graphs, and the right panel shows results on geometric hyperbolic random graphs. We measure performance with the Kendall τ correlation between the ranking obtained from the ground truth and the inferred transition matrices. Error bars represent the 95% confidence interval. The positive impact of constraints on network inference translates to the node ranking task

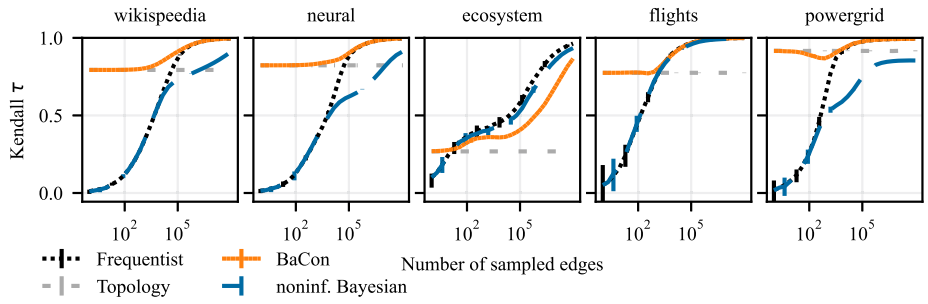

Figure 7 Effects of constraints on the downstream task of node ranking. From left to right, the panels show the results for *wikispeedia*, *neural*, *ecosystem*, *flights*, and *powergrid* networks. We measure performance with the Kendall τ correlation between the ranking obtained from the ground truth and the inferred transition matrices. Error bars represent the 95% confidence interval (mostly too small to be visible). The positive impact of constraints on network inference translates to the node ranking task

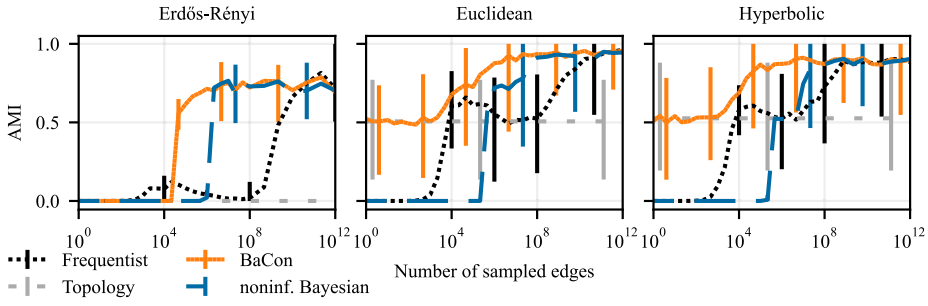

Figure 8 Effects of partial constraints on the downstream task of clustering. The left panel shows results on Erdős Rényi graphs, the middle panel shows results on geometric Euclidean random graphs, and the right panel shows results on geometric hyperbolic random graphs. We measure performance with the Adjusted Mutual Information (AMI) between the clusters obtained from the ground truth and the inferred transition matrices. Error bars represent the 95% confidence interval. The positive impact of constraints on network inference translates to the clustering task

topology, with the only exception of the *powergrid* dataset. However, also in the *powergrid* dataset, BaCon performs better than the frequentist and the noninformative Bayesian approaches.

Cluster detection We present results of experiments on cluster detection in synthetic data in Fig. 8 and Fig. 9, and in empirical data in Fig. 10. These results show that the constraints also improve the detection of clusters. The *flights* dataset might appear as an outlier, where the noninformative Bayesian inference outperforms the other methods. However, this is merely due to the fact that the *flights* dataset exhibits only a single community, which is what the noninformative Bayesian method always predicts a priori. Although the variance of the performances is larger compared to the previous task, we see similar behaviour as in the node ranking experiments. While the frequentist and the noninformative Bayesian approaches have AMI of zero in the limit of small data sizes (indicating no correlation between the identified and ground truth clusters), BaCon defaults to the performance of the unweighted topology. We further inspect this behaviour in Fig. 9, where we show the number of detected clusters in the synthetic graphs introduced in Sect. 4.1. We compare them against the number of clusters detected in the ground truth transition

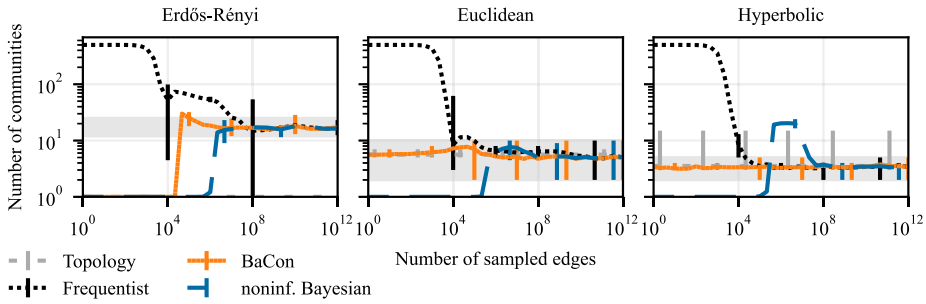

Figure 9 Number of detected clusters by each method for a given number of sampled edges. The left panel shows results on Erdős Rényi graphs, the middle panel shows results on geometric Euclidean random graphs, and the right panel shows results on geometric hyperbolic random graphs. We show the number of communities detected by each method against the horizontal band indicating the interval between 2.5-th and the 97.5-th percentiles of the ground truth number of communities (shaded gray area). Error bars represent the 95% confidence interval. The positive impact of constraints on network inference translates to the clustering task

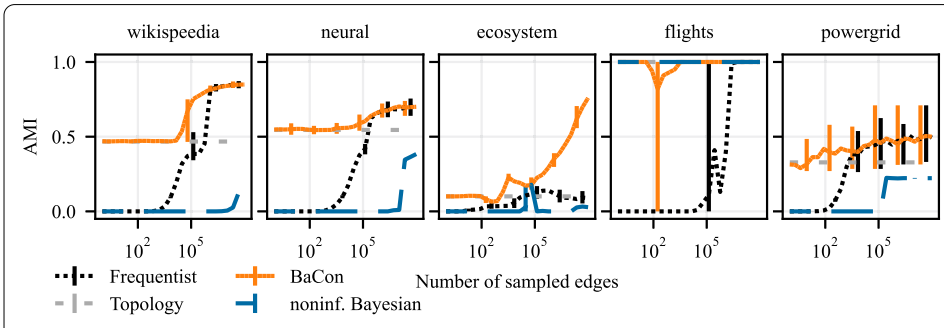

Figure 10 Effects of constraints on the downstream task of clustering. From left to right, the panels show the results for *wikispeedia*, *neural*, *ecosystem*, *flights*, and *powergrid* networks. We measure performance with the Adjusted Mutual Information (AMI) between the clusters obtained from the ground truth and the inferred transition matrices. Error bars represent the 95% confidence interval. The positive impact of constraints on network inference translates to the clustering task

matrix, indicated with the gray horizontal band representing the 95% confidence interval. The noninformative Bayesian method needs considerably more data to detect more than one cluster, while the frequentist method generally overestimated the number of clusters. This means that, although their performance in terms of AMI is similar, they err differently.

Discussion With the experiments presented above, we investigated the consequences of using topological constraints on downstream network analysis tasks, i.e. research question Q3 from Sect. 1. The experiments show that the inclusion of topological constraints in the inference generally improves the performance of downstream network analysis tasks. The results also show that the underfitting of the noninformative Bayesian method and the overfitting of the frequentist method carry over to downstream tasks. Not surprisingly, both methods initially require a sufficient number of observed edges to reach the performance obtained by the unweighted topology alone. This underlines the dilemma that practitioners face when they are confronted with a dataset of edge observations and information of the topology: Which of the two sources will lead to a more reliable analysis

of the underlying network? Instead of choosing between the two sources of information, BaCon allows to include both. As a result, the downstream analysis is as good as for the unweighted topology in the limit of small number of observation, and generally at least as good as the other two methods.

The two outliers, the *ecosystem* and the *powergrid* datasets, require special consideration. Although the inclusion of constraints improved the inference and the performance of the downstream task of cluster detection, they showed different behaviour in the node ranking task. In the *powergrid* dataset, although we see that BaCon outperforms the frequentist and noninformative Bayesian methods, which means that the inclusion of topological constraints improves the inference, we observe that in some cases the unweighted topology leads to better node ranking than the transition matrix inferred with BaCon. In other words, for some data sizes, including observations deteriorates the ranking. The *powergrid* dataset represents an outlier because most of the weights have the same value since they represent voltages. For such an equiprobable distribution of the weights, the uniform distribution ($\alpha = 1$) is not a good prior, because it leads to weights that are more biased than the target. In other words, random fluctuations observed in edge samples are interpreted as actual patterns, which can be viewed as a case of overfitting to the edge observations in this special example. Although BaCon still performs better than the methods that do not use constraints, in Sect. 7, we show how we can further improve the inference by taking prior knowledge about the peakness of the distribution into account.

The *ecosystem* dataset presents the other extreme. This network has a highly biased weight distribution, with ten orders of magnitude between the largest and smallest weights and a Fisher-Pearson coefficient of skewness of 13 (Kokoska and Zwillinger [15]). The flat prior weight distribution (cf. Sect. 3) used in the experiments outlined above makes such biased distributions very unlikely and leads to weights that are less biased than the target. In other words, we are underfitting the observations. This provides a possible explanation why the frequentist method outperforms BaCon, since the frequentist method exhibits a tendency to overfit the observations. However, in the *ecosystem* dataset we also observe that the noninformative Bayesian method leads to better node rankings than BaCon, even though we know from Sect. 5 that it underfits the transition probabilities more than BaCon. The flat prior weight distribution causes underfitting of both BaCon and noninformative Bayesian methods and thus cannot explain alone the better ranking of the noninformative Bayesian method. Therefore, to understand this, we have to consider how the node ranking is impacted by the flat prior weight distribution in conjunction with the topological constraints. The noninformative Bayesian method underfits with the prior of a fully connected network, by which all nodes have equal PageRank scores and an undefined ranking. BaCon underfits with the prior of the unweighted topology; in the topology, nodes have different PageRank scores and their rankings are positively correlated with, but not identical to the ground truth rankings. Because of the underfitting, the scores of nodes remain close to the ones given by the prior, and therefore BaCon outperforms the noninformative Bayesian method in the regime of very small data. However, once we consider a larger number of observed edges, BaCon's prior actually becomes a hindrance. Since the PageRank scores of nodes in the topological prior are not all equal, they require more data to change sufficiently to improve the incorrect rankings. In contrast, the PageRank scores of nodes with the fully connected prior of the noninformative Bayesian method are all equal and thus need less data to change sufficiently to adjust the undefined ranking to the

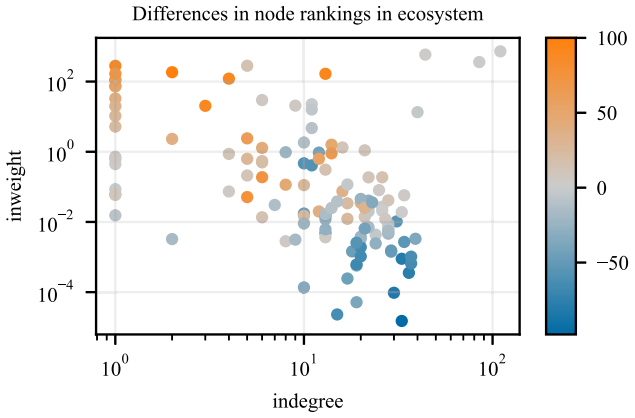

Figure 11 Differences in node ranking in the ecosystem network as a function of indegree and inweight. The color encodes the difference between the ranking nodes get based on the ground truth transition matrix $\mathbf{T}_{gt}$ and the ranking obtained from the unweighted graph topology G. Orange nodes are ranked high by the ground truth and low by the unweighted topology and the blue the opposite. Grey nodes have similar rankings with both matrices. The nodes with high (low) indegree but low (high) inweight are ranked opposite in the topological and weight information and are the reason why BaCon under-performed in the *ecosystem* network (see Fig. 7)

ground truth one. We can peek into the mechanism of how inclusion of the topological prior hurts BaCon's node ranking in the *ecosystem* dataset by looking at Fig. 11. In this figure, we plot the indegree and inweight properties of the nodes, and the discrepancies in their rankings. The high skewness of the weight distribution makes it possible that some nodes have high (low) indegree (x-axis) in the topology but low (high) inweight (y-axis) in the ground truth weighted graph. It is exactly those nodes that have the highest differences between their ground truth rank and their rank from the topology (the difference is encoded in the color of the scatter plot markers). This situation cannot occur with the noninformative Bayesian because all nodes have the same indegree in the prior (the fully connected graph). In Sect. 7, we will show how to incorporate prior knowledge of the skeweness of the weight distribution in the inference step.

Finally, we discuss threats to validity. When the information of constraints G is partial and biased, the inclusion of constraints in the inference can negatively impact the downstream tasks. As an example, take a weighted network in which a node v has the highest centrality and is not connected to all other nodes. If we only know the constraints on the node v, then the resulting (partially known) topological constraint will be a fully connected network except for the connections to and from node v. Thus, looking only at this topology, we would assume that the node v is the *least* important node in the network, and, since BaCon mixes the information present in constraints and in the observations, it would negatively impact node ranking. A similar situation can occur with cluster detection in a weighted network with ground truth clusters C. If the knowledge of the topology of the constraint is biased in such a way that there is a different node grouping C' and we only know constraints across clusters C', then the inferred clusters could also be biased towards C'. In Sect. 7, we will discuss how our approach allows a practitioner to incorporate the information about biases of the two datasources in the inference.

7 Tuning the shape parameter of edge weights

In the previous section, BaCon's results on *ecosystem* showed a subpar performance for node ranking and clustering. Additionally, on *powergrid*, we observed overfitting in the medium samples regime. Both observations relate to the characteristics of the datasets' edge weight distributions: highly biased in *ecosystem* and equiprobable for *powergrid*. In fact, prior knowledge on those systems may lead us to expect that the edge weight distributions exhibit such characteristics: Ecosystems heavily rely on simpler species to introduce energy into the system (Chapin et al. [6]). In a power grid, possible voltages are determined by engineering standards. Leveraging the Bayesian approach of BaCon, we can use these expectations to formulate non-uniform priors for the edge weight distribution. As discussed in Sect. 3, by changing α we obtain prior distributions with different shapes: $\alpha < 1$ favours biased distributions while $\alpha > 1$ favours equiprobable distributions. By matching the parameter α to the expected weight distribution in a given system, we can thus include prior information on the edge weight distribution into our inference, which can make it more data-efficient. In this section, we demonstrate the benefit of this procedure in the *ecosystem* and *powergrid* data sets.

We show the results of the experiments on *ecosystem* in Fig. 12 and of the experiments on *powergrid* in Fig. 13. The experiments are the same as the ones in the previous section but, in addition to the uniform prior ($\alpha = 1$), we now also consider a prior that favours biased $\alpha = 0.01$ and equiprobable distributions $\alpha = 10$. The performance of our method for three different tasks is shown in the three rows of Fig. 12 (top row: inference of transition matrix, middle row: node ranking, bottom row: cluster detection). The columns show the performance for different values of the parameter α (left column: $\alpha = 0.01$, middle column: $\alpha = 1$, right column: $\alpha = 10$). As expected, *ecosystem* performs best with the prior that favours biased distributions and worst with the prior that favours equiprobable dis-

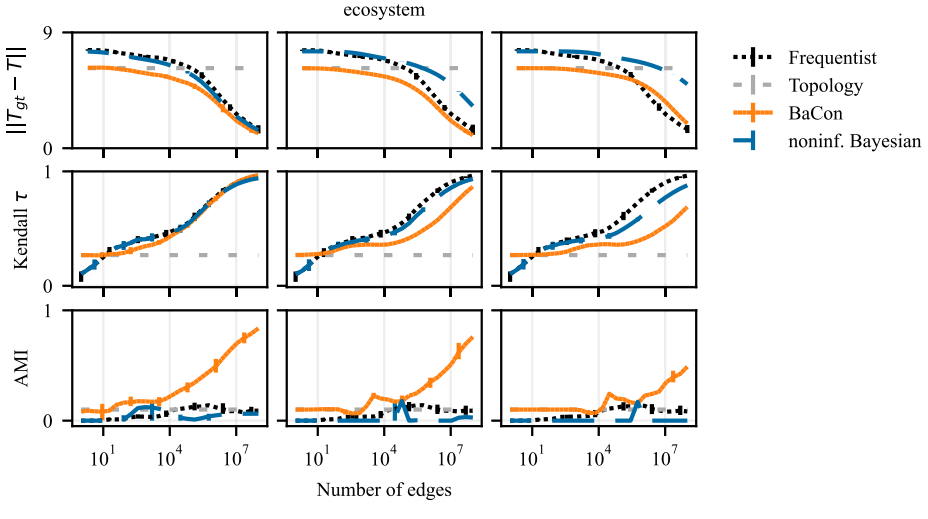

Figure 12 Effect of using different concentration parameters α alongside the topological prior on the *ecosystem* network. Rows indicate the task (top to bottom: transition matrix inference, node ranking, clustering), columns the different values for α (left to right: $\alpha = 0.01$, $\alpha = 1$, $\alpha = 10$). We observe that compared to the value used in the previous experiment ($\alpha = 1$) BaCon performance on the *ecosystem* network improves for the smaller α (left column)

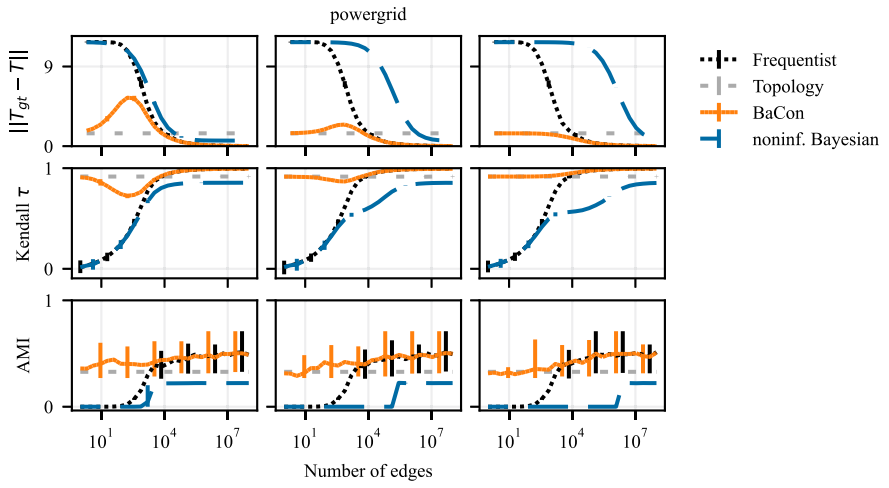

Figure 13 Effect of using different concentration parameters α alongside the topological prior on the *powergrid* network. Rows indicate the task (top to bottom: transition matrix inference, node ranking, clustering), columns the different values for α (left to right: $\alpha = 0.01$, $\alpha = 1$, $\alpha = 10$). We observe that compared to the value used in the previous experiment ($\alpha = 1$) BaCon performance on the *powergrid* network improves for the larger α (right column)

tributions. Conversely, *powergrid* performs best with the prior that favours equiprobable distributions and worst with the prior that favours biased distributions.

Finally, we address the question of the choice of α with Bayesian model selection (Kass and Raftery [11]) as illustrated at the end Sect. 2. We compare choices of α from 50 evenly spaced points on a log scale from 10^{-2} to 10^2. We present the results in Fig. 14. The first three rows show the results of the experiments on *ecosystem* and *powergrid* datasets for the all three tasks. We see a performance improvement compared to Fig. 7, with an improvement of the pagerank results in *ecosystem* and the disappearing of overfitting on the intermediate range for the clustering task in *powergrid*. The fourth row in Fig. 14 shows the value of α chosen by model selection. For BaCon and the noninformative Bayesian method, we have an initial phase where few edges are observed and the model selection chooses a prior that favors equiprobable distributions. With the increase in observations, the α chosen for the noninformative Bayesian stabilizes around the minimum considered value. On the contrary, for BaCon the value of α stabilizes around 0.1 for the *ecosystem* and 10 for the *powergrid*. This difference illustrates that, thanks to the topological constraint, the model selection on BaCon successfully chooses the prior that is justified by the distribution of the weights of existing edges. It chooses the prior that favors equiprobable distributions for *powergrid* and the prior that favors biased distributions for *ecosystem*. In the noninformative Bayesian approach the model selection favors low shape hyperparameters. This is because many edges cannot be observed, and thus have weight zero, which biases the distribution.

Discussion Our results show that using BaCon with suitable informative priors eliminates both the poor performance on *ecosystem* and the overfitting for medium-size samples in *powergrid*. For simplicity, we performed our experiments using two specific values of α ($\alpha = 0.01$, $\alpha = 10$). Nemenman et al. [22] remark that the choice of the concentration parameter α of the Dirichlet distribution which maximizes the variance of the entropy

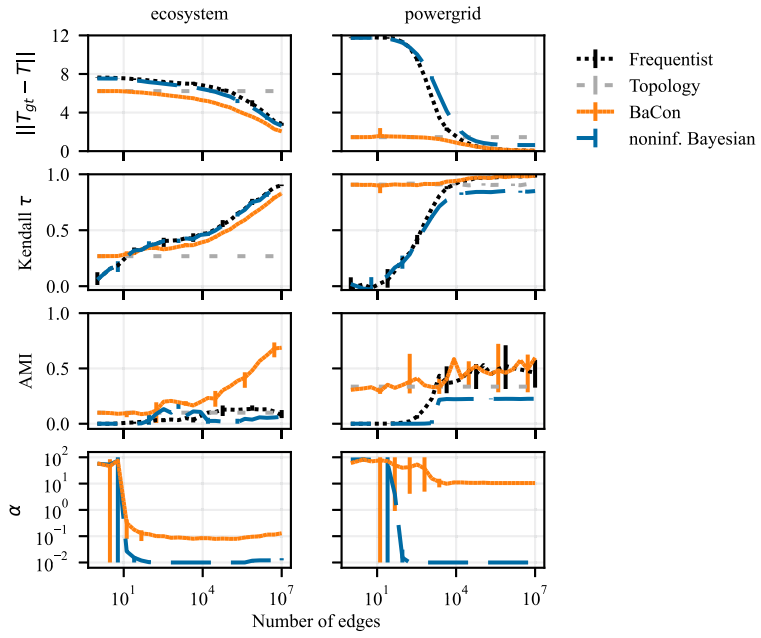

Figure 14 Results of the application of model selection on BaCon and the noninformative Bayesian method with model selection on the *powergrid* and *ecosystem* datasets for the tasks of transition matrix inference (first row), node ranking (second row) and clustering (third row). The fourth row illustrates the prior concentration parameters $\boldsymbol{\alpha}$ resulting from the model selection in the *powergrid* and *ecosystem* datasets for BaCon . The value for BaCon stabilizes around 0.1 for the *ecosystem* and around 10 for the *powergrid*, because *ecosystem* has biased edge weights, and *powergrid* has equiprobable edge weights. Hyperparameter for noninformative Bayesian approach prefers lower values of the hyperparameter – and favors biased distributions –, because some edges never appear

of the multinomial distributions drawn from it is the Schurmann–Grassberger estimator ($\alpha = 1/K$ where K is the alphabet size of the multinomial distribution), albeit they also note that the resulting variance is narrow. Since the skewness and entropy of a multinomial distribution are related, we suspect that the Schurmann-Grassberger estimator might be the optimal choice of a fixed α that makes the least assumptions about the skeweness of the distribution. Finally, we highlight that reducing α would help in the case when the constraint is partially known (and possibly biased), considered at the end of Sect. 6: since spurious edges have zero weights, they effectively increase the skewness of the weight distribution, and thus require lower α. In this case, the reduction of α could also be interpreted as assigning smaller importance to the topological prior when it is less reliable.

Finally, the choice of the free parameter could be solved with a hierarchical Bayesian approach (Kemp et al. [12]). However, this would complicate the inference and may require Monte Carlo simulations, hurting the scalability of the approach. Therefore, we use the model selection described in Sect. 2 and show that its application on the noninformative Bayesian has a tendency to choose priors for biased distribution. In contrast, its application on BaCon identifies priors favoring equiprobable distributions for *powergrid* and biased distributions for *ecosystem*.

8 Related work

In our work, we propose a Bayesian method to infer transition matrices from data capturing a possibly incomplete set of repeated interactions between nodes in a graph that are

subject to topological constraints. The inference of Markov chains' transition matrices from data has rich literature with interesting developments in bioinformatics (Baldi and Brunak [1]) and statistical natural language processing (Manning and Schutze [20]). The problem of inferring the order of higher-order Markov chains has been investigated with connections to statistical mechanics and information-theoretic techniques (Strelioff et al. [37]). The Bayesian estimation of transition probabilities was also used in Peixoto and Rosvall [29] in network and higher-order network models with community structures. While that work is focused on possibilities to cluster nodes and sequences using higher-order patterns, in this paper, we compare the inference methods for the downstream tasks on the inferred first-order networks, and we additionally address the model selection question of making a good choice of shape hyperparameters. The variance of transition probabilities for bigrams and trigrams has been addressed using monograms probabilities to do smoothing on the transition probabilities (MacKay and Peto [19], Manning and Schutze [20]). In contrast to these works, the focus of this work is to use existing topological constraints, and not smoothing transition probabilities in n-gram models. A related approach has recently been proposed to address the detection of the optimal order of higher-order graphical models for causal paths in temporal networks (Petrovic and Scholtes [31]). While this approach is suitable to address the important problem of model selection in higher-order network models (Lambiotte et al. [17]), the challenge of inference and model selection in first-order weighted graphs, which are abundant in practical network analysis, is largely been overlooked. Also, in addition to Petrovic and Scholtes [31], which focuses on optimal-order detection, our work discusses the selection of the shape parameters, which has not been discussed in Petrovic and Scholtes [31], despite it also being important for the higher-order inference. We further explicitly evaluate the accuracy of the inference in incomplete data and assess its impact on downstream network analysis tasks for several synthetic and empirical data sets.

Clearly, the inference of transition matrices from repeated interactions is only one particular challenge that fits into a larger body of works addressing other types of network inference problems, i.e. the inference of weighted and unweighted graphs in noisy or incomplete data, or in data with spurious interactions. A number of works have implicitly addressed the inference of weighted graphs by adapting downstream network analysis tasks to incomplete data, or to data that are subject to errors. In Smiljanić et al. [36] a Bayesian prior is used to avoid overfitting when detecting clusters in data that are subject to both missing or spurious interactions. This prior regularizes the community detection results but, differently from the one considered in this paper, does not use a known network topology as a constraint. Similarly, the issue of network inference has been implicitly addressed with the framework of Graph Neural Networks (GNN). In these works, the inferred network structure and a downstream learning task are jointly optimized by the GNN (Wang et al. [41], Franceschi et al. [8], Wei Jin et al. [10], Zhang et al. [46]). Because of this joint optimization, these works do not aim at a principled inference of an optimal weighted graph model given observations, but rather at finding a network model that optimizes the specific learning task at hand.

Another body of works focuses on inferring the *unweighted topology* of networks from incomplete or biased observations of interactions. Early works tackle the problem from the perspective of a specific domain like, e.g., social systems (Butts [4]) or connectomes in neuroscience (Priebe et al. [33]) (further references can be found in Young et al. [43]). More

recent works define the problem in more general terms as the inference of both topology and a *data model* connecting the topology to the observations. Here, a *data model* is defined as a model that produces observations as a function of the network topology and additional parameters. In Newman [23], the network topology is obtained through a procedure of expectation-maximization from potentially incomplete or noisy data. In Casiraghi et al. [5], the problem is tackled from the perspective of the statistical significance of the interactions. The method provides a way to infer statistically significant edges from edges that could have happened at random in the system.

Methods similar to the ones above, but using a Bayesian approach for the inference of weighted graphs, were proposed where the posterior is obtained through a Markov Chain Monte Carlo procedure (Young et al. [43], Peixoto [26]). In Peixoto [27], the prior is a generative model that couples the network inference with the detection of communities. In Rabbat et al. [34] the directedness of the network is inferred from undirected co-occurrence data. The data are assumed to be generated by random walks, and Expectation Maximization is used to estimate transition probabilities.

Unlike the methods above, we assume that we have (at least partial) knowledge on which interactions are possible. We further consider situations where no spurious interactions can be observed, i.e. we have access to data that represent a subset of the possible interactions. We use this knowledge to define a Bayesian prior that, differently from other Bayesian priors in the literature, explicitly constrains which interactions are possible. We obtain a weighted graph inference method that can be solved analytically, thus being computationally efficient. Different from existing methods, we also show how the parameters controlling our prior can be tuned depending on the expected skewness of the underlying weight distribution, thus further improving the data efficiency of our method.

9 Conclusion

In this work, we address the problem of inferring transition matrices from data capturing an incomplete sample of repeated interactions in networked systems with a known topology. Examples include data on users navigating information networks with a known hyperlink structure, passengers travelling in a transportation network with known physical topology, or observed social interactions in systems with known social structure. To the best of our knowledge, no existing inference method has specifically addressed this problem, despite its large practical relevance for network analysis tasks that rely on transition matrices constructed from weighted graphs.

We address this issue with a Bayesian approach, where the prior distribution captures our knowledge of the network topology. An experimental evaluation of our method in synthetic and empirical datasets shows that it considerably outperforms a common frequentist inference method and a noninformative Bayesian approach both in terms of the accuracy of the inferred transition matrix, and in terms of the results of downstream network analysis tasks. The prior based on topological constraints regularizes the inferred probabilities and thus prevents overfitting in small data sets (exhibited by the frequentist method). It simultaneously limits the degrees of freedom, which prevents underfitting (exhibited by the noninformative Bayesian approach). Our results show that such a prior is effective even when the knowledge of the constraints is partial. Moreover, thanks to its analytical tractability, our approach does not require expensive simulations. It just requires Bayesian updating based on a simple counting of interaction occurrences. Highlighting

issues in networks with biased or peaked distributions of edge weights, we finally show how the adjustment of a parameter that controls the prior distribution of edge weights can be used to further improve the data efficiency of our method.

In summary, we propose an intuitive and elegant method for a common problem in network analysis. Referring to the adjustment of the prior to the expected edge probability distribution, in future works we seek to address the question how a suitable choice of this parameter can be learned from the data. We further expect that our method can be used to improve network inference in situations where we do *not* have access to a network constraint, e.g. by combining it with existing methods to infer the *unweighted topology*. Finally, considering the fact that our method is particularly useful in situations in which the amount of data is small compared to the dimension of the space of possible interactions (i.e. the degrees of freedom determined by the network topology), we expect it to be of considerable interest for the growing community addressing inference tasks in higher-order network models (Lambiotte et al. [17]).

Acknowledgements
Vincenzo Perri, Luka Petrović, and Ingo Scholtes acknowledge support by the Swiss National Science Foundation, grant 176938.

Abbreviations
MLE, Maximum Likelihood Estimation; AMI, Adjusted Mutual Information; GNN, Graph Neural Network.

Availability of data and materials
The code and the results of the current study are available in the Zenodo repository (Zenodo [44]). All real world data sets used in this manuscript are publicly available (Verasztó et al. [40], Ulanowicz and DeAngelis [39], West and Leskovec [42], TransStat [38], Kim et al. [13], Peixoto [28]) and are also collected in a dedicated Zenodo repository (ZenodoData [45]).

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
Concept: LP. Mathematical foundations: LP. Software: LP and VP. Experiments: LP and VP. Draft manuscript: LP and VP. Final manuscript: LP, VP and IS. Supervision and project management: IS. All authors read and approved the final manuscript.

Author details
[1] Data Analytics Group, Department of Informatics, University of Zurich, Binzmühlestrasse 14, CH-8050 Zurich, Switzerland.
[2] Chair of Machine Learning for Complex Networks, Center for Artificial Intelligence and Data Science, Julius-Maximilians-Universität Würzburg, John-Skilton-Strasse 8a, 97074 Würzburg, Germany.

Received: 22 August 2022 Accepted: 28 August 2023 Published online: 11 October 2023

References

1. Baldi P, Brunak S (2001) Bioinformatics: the machine learning approach. MIT Press, Cambridge
2. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: structure and dynamics. Phys Rep 424(4–5):175–308
3. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P (2017) Geometric deep learning: going beyond Euclidean data. IEEE Signal Process Mag 34(4):18–42
4. Butts CT (2003) Network inference, error, and informant (in) accuracy: a Bayesian approach. Soc Netw 25(2):103–140
5. Casiraghi G, Nanumyan V, Scholtes I, Schweitzer F (2017) From relational data to graphs: inferring significant links using generalized hypergeometric ensembles. In: International conference on social informatics. Springer, Berlin, pp 111–120
6. Chapin FS, Matson PA, Mooney HA, Vitousek PM (2002) Principles of terrestrial ecosystem ecology. Springer, Berlin
7. Chung F (2007) The heat kernel as the pagerank of a graph. Proc Natl Acad Sci USA 104(50):19735–19740
8. Franceschi L, Niepert M, Pontil M, He X (2019) Learning discrete structures for graph neural networks. In: International conference on machine learning. PMLR, pp 1972–1982
9. Grover A, Leskovec J (2016) node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 855–864
10. Jin W, Ma Y, Liu X, Tang X, Wang S, Tang J (2020) Graph structure learning for robust graph neural networks. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp 66–74
11. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90(430):773–795

12. Kemp C, Perfors A, Tenenbaum JB (2007) Learning overhypotheses with hierarchical Bayesian models. Dev Sci 10(3):307–321
13. Kim H, Olave-Rojas D, Álvarez-Miranda E, Son S-W (2018) In-depth data on the network structure and hourly activity of the central Chilean power grid. Sci Data 5(1):1–10
14. Kloster K, Gleich DF (2014) Heat kernel based community detection. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1386–1395
15. Kokoska S, Zwillinger D (2000) CRC standard probability and statistics tables and formulae. CRC Press, Boca Raton
16. Krioukov D, Papadopoulos F, Kitsak M, Vahdat A, Boguná M (2010) Hyperbolic geometry of complex networks. Phys Rev E 82(3):036106
17. Lambiotte R, Rosvall M, Scholtes I (2019) From networks to optimal higher-order models of complex systems. Nat Phys 1
18. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019–1031
19. MacKay DJ, Peto LCB (1995) A hierarchical Dirichlet language model. Nat Lang Eng 1(3):289–308
20. Manning C, Schutze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge
21. Masuda N, Porter MA, Lambiotte R (2017) Random walks and diffusion on networks. Phys Rep 716:1–58
22. Nemenman I, Shafee F, Bialek W (2001) Entropy and inference, revisited. Adv Neural Inf Process Syst 14
23. Newman MEJ (2018) Network structure from rich but noisy data. Nat Phys 14(6):542–545
24. Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web. Technical report, Stanford InfoLab
25. Peel L, Peixoto TP, De Domenico M (2022) Statistical inference links data and theory in network science. Nat Commun 13(1):1–15
26. Peixoto TP (2018) Reconstructing networks with unknown and heterogeneous errors. Phys Rev X 8(4):041011
27. Peixoto TP (2019) Network reconstruction and community detection from dynamics. Phys Rev Lett 123(12):128301
28. Peixoto TP (2020) The netzschleuder network catalogue and repository. https://networks.skewed.de/
29. Peixoto TP, Rosvall M (2017) Modelling sequences and temporal networks with dynamic community structures. Nat Commun 8(1):582
30. Perozzi B, Al-Rfou R, Deepwalk SS (2014) Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 701–710
31. Petrovic LV, Scholtes I (2022) Learning the Markov order of paths in graphs. In: Proceedings of the ACM web conference 2022, pp 1559–1569
32. Pons P, Latapy M (2006) Computing communities in large networks using random walks. J Graph Algorithms Appl Citeseer
33. Priebe CE, Sussman DL, Tang M, Vogelstein JT (2015) Statistical inference on errorfully observed graphs. J Comput Graph Stat 24(4):930–953
34. Rabbat MG, Figueiredo MA, Nowak RD (2008) Network inference from co-occurrences. IEEE Trans Inf Theory 54(9):4053–4068
35. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105(4):1118–1123
36. Smiljanić J, Edler D, Rosvall M (2020) Mapping flows on sparse networks with missing links. Phys Rev E 102(1):012302
37. Strelioff CC, Crutchfield JP, Hübler AW (2007) Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. Phys Rev E 76(1):011106
38. TransStat RITA (2014) Origin and destination survey database. http://www.transtats.bts.gov/Tables.asp?DB_ID=125
39. Ulanowicz RE, DeAngelis DL (2005) Network analysis of trophic dynamics in south Florida ecosystems. US Geological Survey Program on the South Florida Ecosystem 114:45
40. Verasztó C, Jasek S, Gühmann M, Shahidi R, Ueda N, Beard JD, Mendes S, Heinz K, Bezares-Calderón LA, Williams E et al (2020) Whole-animal connectome and cell-type complement of the three-segmented platynereis dumerilii larva. bioRxiv
41. Wang R, Mou S, Wang X, Xiao W, Ju Q, Shi C, Xie X (2021) Graph structure estimation neural networks. In: Proceedings of the web conference 2021, pp 342–353
42. West R, Leskovec J (2012) Human wayfinding in information networks. In: Proceedings of the 21st international conference on world wide web, pp 619–628
43. Young J-G, Cantwell GT, Newman MEJ (2020) Bayesian inference of network structure from unreliable data. J Complex Netw 8(6):cnaa046
44. Zenodo (2022) Reproducibility package. https://doi.org/10.5281/zenodo.7007374
45. ZenodoData (2022) Data availability package. https://doi.org/10.5281/zenodo.8121356
46. Zhang Y, Pal S, Coates M, Ustebay D (2019) Bayesian graph convolutional neural networks for semi-supervised classification. In: The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, the ninth AAAI symposium on educational advances in artificial intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, Menlo Park, pp 5829–5836. https://doi.org/10.1609/aaai.v33i01.33015829

Publisher's Note