When is a Network a Network? Multi-Order Graphical Model Selection in Pathways and Temporal Networks

Ingo Scholtes

ETH Zürich
CH-8092, Zürich, Switzerland
ischoltes@ethz.ch

ABSTRACT

We introduce a framework for the modeling of sequential data capturing *pathways* of varying lengths observed in a network. Such data are important, e.g., when studying click streams in the Web, travel patterns in transportation systems, information cascades in social networks, biological pathways, or time-stamped social interactions. While it is common to apply graph analytics and network analysis to such data, recent works have shown that temporal correlations can invalidate the results of such methods. This raises a fundamental question: When is a network abstraction of sequential data justified? Addressing this open question, we propose a framework that combines Markov chains of multiple, higher orders into a multi-layer graphical model that captures temporal correlations in pathways at multiple length scales simultaneously. We develop a model selection technique to infer the optimal number of layers of such a model and show that it outperforms baseline Markov order detection techniques. An application to eight real-world data sets on pathways and temporal networks shows that it allows to infer graphical models that capture both topological and temporal characteristics of such data. Our work highlights fallacies of network abstractions and provides a principled answer to the open question when they are justified. Generalizing network representations to multi-order graphical models, it opens perspectives for new data mining and knowledge discovery algorithms.

1 INTRODUCTION

The modeling and analysis of sequential data is an important task in data mining and knowledge discovery, with applications in text mining, click stream analysis, bioinformatics and social network analysis. An interesting class of data relevant in these contexts are those that provide us with collections of observed *pathways*, i.e. multiple (typically short) sequences of *vertices* traversed by paths in an underlying graph or network. Examples include traces of information propagating in (online) social networks, click streams of users in hyperlinked documents, biochemical cascades in biological signaling networks, or contact sequences emerging from time-stamped data on social interactions.

The graph topology underlying these systems has enticed researchers and practitioners to apply *graph analytics* and *network analysis*, e.g., to make statements about node centralities, cluster and community structures, or subgraph and motif patterns. While these methods have their merits, recent works have voiced concerns about their naive application to complex data [2, 43]. In particular, network-analytic methods make the fundamental assumption that *paths are transitive*, i.e. that the existence of paths from a to b and from b to c implies a *transitive path* from a via b to c. As shown recently, non-trivial temporal correlations in pathways and temporal networks can invalidate this assumption [15, 20]. As a result, network-based modeling and mining techniques yield wrong results, e.g., about cluster structures, the ranking of nodes, or dynamical processes such as information propagation. Addressing this issue, recent works have thus argued for *higher-order network models* that capture both temporal and topological characteristics of sequential data [21, 22, 26, 27, 38, 41].

Contributions Going beyond these prior works, we advance the state-of-the-art in sequential data mining as follows: (1) We introduce a *multi-order graphical modeling framework* tailored to data capturing multiple variable-length pathways in networks. Our approach combines multiple higher-order Markov models into a multi-layer model consisting of *De Bruijn graphs* with multiple dimensions. Different from previous approaches, this allows us to capture temporal correlations with multiple correlation lengths simultaneously. (2) We introduce a model selection technique that accounts for the structure of pathway data and for topological constraints imposed by the underlying graph that were neglected in prior works. Using synthetic and real-world data, we show that this approach dramatically improves the modeling of pathways and temporal networks, opening new perspectives for the analysis of click streams, biological pathways and time-stamped social networks. (3) Using PageRank as a case study, we show that correlations in sequential data can invalidate the application of graph-analytic methods. We finally demonstrate that our framework allows to generalize such methods to higher-order models that capture both topological and temporal patterns in a simple, static representation.

Our work not only challenges naive applications of network-analytic methods to sequential data. It also provides a principled method to (i) decide *when* a network abstraction of such data is justified, and (ii) infer *optimal* higher-order graphical models that can be used to generalize network analysis techniques.

2 RELATED WORK

The analysis of sequential data has important applications in areas like natural language processing, data compression, behavioral modeling or bioinformatics [6, 13, 42]. Considering the focus of this paper, here we limit our review of the relevant literature to works addressing the modeling of (i) sequential data on pathways in graphs, or (ii) time-stamped data on temporal or dynamic graphs.

Click streams or *user trails* in the Web are one example for pathway data, with important applications in user modeling and information retrieval. A number of recent works have studied Markov chain models of human click paths [4, 14, 23, 29, 35, 37]. Chierichetti et al. [4] study whether the Markovian assumption underlying models that only take into account the topology of the underlying Web graph is justified. They find that accounting for non-Markovian characteristics, which are due to correlations in the ordering of traversed pages, improves

the prediction performance of a variable-order Markov chain model. Similarly, West and Leskovec [35] model navigation paths of users playing the Wikispeedia game, finding that incorporating correlations not captured by the topology of the Wikipedia article graph improves the performance of a target prediction algorithm. Taking a model selection approach, Singer et al. [29] argue that higher-order Markov models are not justified for click stream data at the page level, while they are warranted for coarse-grained data at a topic or category level.

Apart from click streams, the influence of order correlations has also been studied in other types of pathway data such as, e.g., human travel patterns [19, 21, 22, 27], knowledge flow in scientific communication [21], or cargo traces in logistics networks [38]. Like for click streams, it was found that correlations in real data on networked systems do not justify the Markovian assumption implicitly made by typical graph-based modeling techniques. Similar results have been obtained for high-frequency data on *dynamic* or *temporal graphs*, i.e. relational data that capture the detailed timing and ordering of relations. Thanks to improved data collection and sensing technology, such data are of growing importance in various settings. Important applications include, e.g., cluster detection in temporal graphs capturing economic transactions or social interactions [16, 22], ranking nodes in dynamic social networks [26, 40], or identifying frequent interaction patterns in communication networks [39]. Despite their importance, the analysis of such data is still a considerable challenge. In particular, it has been shown that temporal correlations in the sequence of time-stamped interactions shape connectivity, cluster structures, node centralities, and dynamical processes in temporal networks [12, 15, 20, 22]. This questions applications of data mining techniques based on time-aggregated or time-slice abstractions, which neglect the ordering of interactions.

In summary, these works show that autocorrelations in pathways and temporal networks hinder topology-based modeling techniques, with important consequences for sequential pattern mining and graph analytics. *Higher-order network modeling techniques*, which build on higher- or variable-order Markov models, have been proposed to address this problem [19, 21, 22, 26, 27, 38]. While there is agreement about the need for such techniques, principled methods to decide (i) when the use of network-based methods is invalid and (ii) which higher-order model should be used for a given data set were investigated only recently [19, 29]. Moreover, existing works have mostly focused on modeling techniques that account for temporal correlations at a single fixed length, while real-world sequential data are likely to exhibit multiple correlation lengths simultaneously. Finally, using state-of-the-art Markov chain inference techniques, previous works did not account for special characteristics of data on multiple, independent paths with varying lengths that are observed in a known graph topology. Proposing a model selection technique tailored to such sequential data, this paper addresses this research gap. Interpreting time-stamped data on temporal networks as one possible source of pathway data that can be modeled with our framework, we further highlight interesting and previously unknown relations between problems addressed in sequence modeling, pattern mining and (dynamic) graph analysis.

3 PRELIMINARIES

We first introduce the problem addressed in our work and provide some preliminaries on (higher-order) Markov chain models of pathway data. Assume we are given a multi-set $S = \{p_1, \ldots, p_N\}$ with N independent observations of sequences p_i, representing *paths* of varying lengths $l_i \geq 0$ in a graph $G = (V, E)$ with vertices V and (directed) edges $E \subseteq V \times V$. Each of these paths $p_i = (v_0 \to v_1 \to \ldots \to v_{l_i})$ is an ordered tuple of $l_i + 1$ vertices such that $(v_i, v_{i+1}) \in E$ for all $i \in [0, l_i - 1]$. The length l_i of path p_i is the number of edges that it traverses, i.e. a (trivial) path $p = (v_0)$ consisting of a single vertex has length zero. Depending on the context, S could capture click paths of users in the Web, chains of molecular interactions in a cell or itineraries of passengers in a transportation network. We further assume that the underlying graph G captures topological constraints such as, e.g., hyperlinks between Web documents influencing click paths, molecular structures limiting possible reactions, or possible routes in a transportation network.

Interpreting vertices as *categories*, we can view paths as *categorical sequences* and we can consider a probabilistic model that provides a probability $P(S)$ to observe a given multi-set S. Higher-order Markov chains are a powerful class of probabilistic models, with applications in data analysis, inference and prediction tasks [1, 31]. Considering paths as multiple sequences of random variables, we can define a discrete time Markov chain of order k over a discrete state space V that assigns probabilities to each consecutive vertex. For this, we assume that the Markov property holds, i.e. for each v_i

$$P(v_i | v_0 \to \ldots \to v_{i-1}) = P(v_i | v_{i-k} \to \ldots \to v_{i-1}) \tag{1}$$

where k is the "memory" of the model. I.e., the i-th vertex on a path depends (only) on the k previously traversed vertices.

We call $P^{(k)} := P(v_i | v_{i-k} \to \ldots \to v_{i-1})$ the transition probability of a k-th order Markov chain. It probabilistically generates sequences by means of repeated transitions between vertices, each extending a sequence by a single vertex depending on the k previous vertices. For $k = 0$ we obtain transition probabilities $P^{(0)}(v_i)$, i.e., each step v_i is independent of previous steps. Importantly, the independence assumption of such a *zero-order model* does not allow us to selectively generate paths constrained to a given graph, since *any* sequence of vertices with non-zero probabilities can be generated, independent of whether it corresponds to a path in the underlying graph or not. For $k = 1$, the model keeps a memory of one step, i.e., the probability $P^{(1)}(v_i | v_{i-1})$ to "move" to vertex v_i depends on the "current" vertex v_{i-1}. The dyadic dependencies captured in such a *first-order model* allow us to assign zero probabilities $P^{(1)}(v_i | v_{i-1}) = 0$ to those transitions for which no corresponding edge exists, i.e. $(v_{i-1}, v_i) \notin E$. Hence, first-order models are the simplest models able to generate paths constrained to a graph. For $k > 1$, a *k-th order model* can additionally capture higher-order dependencies, i.e. correlations in the sequence of vertices that go beyond topological constraints imposed by the underlying graph.

An important (and non-trivial) question in the study of categorical sequence data is which order k of a Markov chain is needed to model (or summarize) a given data set. It naturally relates to prediction and compression tasks and has received attention from researchers in data mining, signal processing and statistical inference. Specifically, higher-order Markov chain models provide a foundation for (Bayesian) model selection and inference techniques that are based on the likelihood function [1]. For a given transition probability $P^{(k)}$ of a k-th order model M_k, the likelihood $L(M_k | p)$ under an observed path $p = (v_0 \to$

$\ldots \rightarrow v_l)$ is given as:

$$L(M_k|v_0 \rightarrow \ldots \rightarrow v_l) = \prod_{i=k}^{l} P^{(k)}(v_i|v_{i-k} \rightarrow \ldots \rightarrow v_{i-1}) \quad (2)$$

For our scenario of a multi-set set S of (statistically independent) paths, the likelihood of a k-th order model M_k is then

$$L(M_k|S) = \prod_{j=1}^{N} L(M_k|p_j) \quad (3)$$

where p_j is the j-th observed path in S. This allows us to perform a maximum likelihood estimation (MLE) of transition probabilities $\hat{P}^{(k)}$ for any order k based on a set of observed pathways S. In other words, we can "learn" the parameters of a k-th order graphical model based on the frequencies of paths in a data set. To formally define this, we first introduce the notion of a *sub path*. For two paths $p = (p_0 \rightarrow \ldots \rightarrow p_k)$ and $(q = q_0 \rightarrow \ldots \rightarrow q_l)$ with $k \leq l$, we say that p is sub path of q with length k ($p \sqsubseteq q$) iff $\exists a \geq 0$ such that $q^{i+a} = p^i$ for $i \in [0, k]$. In other words: $p \sqsubseteq q$ iff path p occurs in (or is equal to) path q. With this, the transition probabilities $\hat{P}^{(k)}$ of a k-th order model that maximize likelihood can be calculated as

$$\hat{P}^{(k)}(v_i|v_{i-k} \ldots \rightarrow v_{i-1}) = \frac{|\{(v_{i-k} \ldots \rightarrow v_i) \in S_k\}|}{\sum_{w \in V} |\{(v_{i-k} \ldots \rightarrow v_{i-1} \rightarrow w) \in S_k\}|} \quad (4)$$

where S_k is the multi-set of *sub paths of length k* of S, i.e. we define $S_k := \{p \in V^k : \exists q \in S : p \sqsubseteq q\}$. Hence, we infer the transition probabilities of a k-th order Markov chain based on the relative frequencies of *sub paths* of length k.

We conclude this section by commenting on the relation between higher-order Markov chains and graph abstractions of pathway data. For $k = 1$, inferred probabilities $\hat{P}^{(1)}$ capture relative frequencies of traversed *edges* (i.e. sub paths of length one) in the graph. Such a first-order model is given by a *weighted graph*, where edges capture the topology and weights capture relative frequencies at which paths traverse edges. For $k > 1$, transition probabilities are calculated based on relative frequencies of *longer* paths, capturing correlations in sequences of vertices that are not due to the graph topology. Such *higher-order models* can be visualized by a construction that resembles high-dimensional *De Bruijn graphs* [5]. It is based on the common representation of Markov chains of order k on state space V as *first-order* Markov chains on an extended state space V^k. Each transition $P(v_i|v_{i-k} \rightarrow \ldots \rightarrow v_{i-1})$ that corresponds to a path of length k is represented by an edge between two *k-th order vertices* $(v_{i-k}, \ldots, v_{i-1})$ and $(v_{i-k+1}, \ldots, v_i)$ in an extended state space V^k. The "memory" of length k is then encoded by higher-order vertices and each transition shifts it by one vertex.

This provides graphical models $G^{(k)}$ for different orders k, where the topology of the first-order model $G = G^{(1)}$ corresponds to commonly used network abstractions. For $k > 1$ we obtain *higher-order graphical models* $G^{(k)}$, which represent both the topology of the graph as well as correlations in the sequence of vertices not captured by G [27]. A k-th order graphical model particularly encodes deviations from the path transitivity assumption that result from the statistics of (sub) paths of length k, while its graphical interpretation corresponds to the assumption that paths *longer* than k are transitive. Hence, k-th order graphs $G^{(k)}$ can be seen as natural generalization of network abstractions for sequential data. They account for correlations that invalidate the transitivity assumption made by a first-order model.

$S = \{(B \rightarrow D), (B \rightarrow C),$
$(D \rightarrow A), (D \rightarrow B), (A \rightarrow B)$
$(B \rightarrow C \rightarrow A), (A \rightarrow B \rightarrow D)$
$(D \rightarrow A \rightarrow B), (B \rightarrow D \rightarrow B)$
$(C \rightarrow A \rightarrow B), (D \rightarrow B \rightarrow D)$
$(B \rightarrow D \rightarrow A), (A \rightarrow B \rightarrow C)$
$(B \rightarrow D \rightarrow B \rightarrow D)$
$(D \rightarrow A \rightarrow B \rightarrow D)$
$(A \rightarrow B \rightarrow C \rightarrow A)$
$(A \rightarrow B \rightarrow D \rightarrow B \rightarrow D)$
$(D \rightarrow B \rightarrow D \rightarrow B \rightarrow D)$
$(C \rightarrow A \rightarrow B \rightarrow D \rightarrow B \rightarrow D)$
$(B \rightarrow D \rightarrow B \rightarrow D \rightarrow B \rightarrow D), \ldots\}$

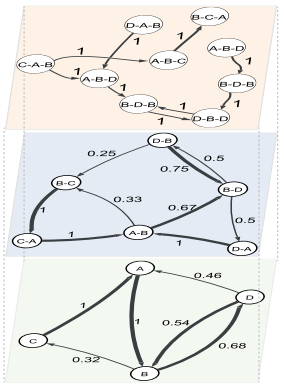


Figure 1: Example for three layers of (higher-order) graphical models (right) for toy example S of paths (left) in a graph with vertices $V = \{A, B, C, D, E\}$ connected by six edges ($G^{(1)}$).

Fig. 1 shows an illustrative example for a multi-set S of paths (left) and the corresponding higher-order graphical models $G^{(k)}$ for different orders $k \geq 1$.

4 MULTI-ORDER GRAPHICAL MODELS

We now introduce the multi-order graphical modeling framework that constitutes the main contribution of our work. It relies on the capability of higher-order models to capture correlations in sequential data that are neglected by common graph or network abstractions. Going beyond previous works, we (i) infer multi-layer graphical models that consider multiple correlation lengths simultaneously and (ii) provide a statistically principled answer to the question which order k of a graphical model $G^{(k)}$ should be used to analyze a given data set.

While it is tempting to address this problem with standard Markov chain inference and order detection techniques, it is important to take into account special characteristics of pathway data. We first observe that the likelihood calculation for a k-th order Markov chain neglects, by construction, the first k vertices on a path (cf. Eq. 2). This is not an issue for a single long sequence. However, it poses problems when modeling large numbers of (typically short) paths. Depending on the distribution of path lengths, the number of paths entering the likelihood calculation in Eq. 3 is likely to decrease as the order k increases, which complicates model selection. This problem is often addressed by concatenating multiple pathways to a single sequence, possibly separated by a delimiter symbol. However, as we show later, this introduces issues that question the use of standard sequence mining techniques.

We address these issues by means of graphical models that combine multiple layers of Markov chain models of multiple orders to a single multi-order model. For this, we first infer multiple k-th order models M_k for $k = 0, \ldots, K$ up to a maximum order K as described in section 3, i.e. we "learn" the parameters of each k-th order model M_k using Eq. 4. We then combine them into a multi-order graphical model $\bar{M}_K$, where each model layer captures correlations in the sequence of vertices at a specific length k. For the resulting model, we then iteratively define the probability $\bar{P}^{(K)}$ to generate a path $(v_0 \rightarrow \ldots \rightarrow v_l)$ of length l based on transition probabilities $P^{(k)}$ of *all* model layers k *up to*

maximum order K as:

$$\bar{P}^{(K)}(v_0 \to \ldots \to v_l) = \prod_{k=0}^{K} P^{(k)}(v_k | v_0 \to \ldots \to v_{k-1}) \prod_{i=K+1}^{l} P^{(K)}(v_i | v_{i-K} \to \ldots \to v_{i-1}) \tag{5}$$

The first product multiplies the transition probabilities $P^{(k)}$ in $K+1$ model layers with increasing order and prefix length $k = 0, \ldots, K$. For paths longer than the maximum order K, the second product additionally accounts for $l - K$ transitions in the layer with the maximum order K. To illustrate this, consider the probability of a path $p = (v_0 \to v_1 \to v_2 \to v_3 \to v_4)$ of length $l = 4$ in a multi-order model with maximum order $K = 2$. From Eq. 5, we get

$$\bar{P}^{(2)}(p) = P^{(0)}(v_0) \cdot P^{(1)}(v_1|v_0) \cdot P^{(2)}(v_2|v_0 \to v_1) \cdot P^{(2)}(v_3|v_1 \to v_2) \cdot P^{(2)}(v_4|v_2 \to v_3)$$

where each of the first three products corresponds to a single transition in the model layer k with increasing order and prefix length k. The last two products are due to two additional transitions in the layer with maximum order $K = 2$ and a prefix length of two.

Based on Eq. 5, we define the likelihood $L(\bar{M}_K)$ of a multi-order model with maximum order K under a set S of observed paths as

$$L(\bar{M}_K|S) = \prod_{j=1}^{N} \bar{P}^{(K)}(p_j) \tag{6}$$

where p_j is the j-th path in S. We can then perform a maximum likelihood estimation analogous to Eq. 4. Here we use sub paths with length *exactly* k to estimate transition probabilities of layers $k < K$, while paths with length longer or equal than K are used to estimate transition probabilities of layer K. We obtain a multi-layer model for paths of varying lengths, each layer being a higher-order graphical model that captures correlations at length k (cf. Fig. 1).

4.1 Detection of optimal maximum order

The modeling framework above allows to develop a method to infer the *optimal* maximum order K_{opt} of a multi-order model for a given set of pathways S. That is, we address the important question how many layers of higher-order graphical models are needed to study a given data set: An optimal maximum order $K_{opt} = 1$ signifies that pathways do not contain correlations that break the transitivity assumption made when using a first-order graphical model. This correspond to situations where the structure (and frequency) of observed paths can be explained based on the underlying (first-order) network. We argue that in this (and only in this) case, the application of network-analytic methods is justified. For data with $K_{opt} > 1$, their application is misleading since order correlations break the transitivity of paths in the first-order network [20, 26, 27]. In other words, for $K_{opt} > 1$ the observed pathways invalidate the assumption of path transitivity implicitly made by standard network-analytic methods. We will show that a generalization of these methods to the higher-order graphs that constitute the layers of our multi-order model provides a simple yet efficient way to analyze data that do not warrant standard network abstractions.

Our method to infer the optimal maximum order of a multi-order model is based on the likelihoods of candidate multi-order models, which combine higher-order models up to different maximum orders K (cf. Eq. 6). Clearly, simply maximizing $L(\bar{M}_{K_{opt}}|S)$ would overfit the data since the inclusion of additional model layers trivially increases the likelihood at the expense of increased model complexity. Applying Occam's razor, we are instead interested in a multi-order graphical model that balances model complexity and explanatory power for the observed set of pathways.

Several techniques to avoid overfitting higher-order Markov chains have been proposed and methods based on the Bayesian or Aikake Information Criterion are frequently used for this purpose [11, 28, 31]. However, previous works have not accounted for special characteristics of pathway data, which is why we introduce a different approach that utilizes the *nested structure* of multi-order models. For this, consider two multi-order models $\bar{M}_K$ and $\bar{M}_{K+1}$, which combine higher-order graphical models up to order K and $K+1$ respectively. We consider the model $\bar{M}_K$ as the *null model*, while $\bar{M}_{K+1}$ provides the alternative model. The likelihood ratio $\frac{L(\bar{M}_K|S)}{L(\bar{M}_{K+1}|S)}$ captures how much more likely S is under the (more complex) model $\bar{M}_{K+1}$ compared to the (simpler) null model $\bar{M}_K$. It also allows to calculate a p-value that can be used to reject the alternative model $\bar{M}_{K+1}$ in favor of model $\bar{M}_K$.

To calculate this p-value we must generally derive the distribution of likelihood ratios, which is possible only in simple cases. We can avoid this by considering that $\bar{M}_K$ and $\bar{M}_{K+1}$ are *nested*, i.e. the model $\bar{M}_K$ is a special case in the parameter space of the more complex model $\bar{M}_{K+1}$. This follows from the fact that probabilities of paths of length $k+1$ in layer $k+1$ can be set to the probabilities resulting from *two* transitions in the layer k. This nestedness allows to apply Wilk's theorem [36], which states that the distribution of likelihood ratios between two nested models $\bar{M}_K$ and $\bar{M}_{K+1}$ asymptotically follows a chi-squared distribution $\chi^2(x)$, where x is the difference in the degrees of freedom between $\bar{M}_{K+1}$ and $\bar{M}_K$. With this, we calculate the p-value of the null hypothesis $\bar{M}_K$ using the cumulative distribution function of the chi-squared distribution as

$$p = 1 - \frac{\gamma\left(\frac{d(K+1)-d(K)}{2}, -\log\frac{L(\bar{M}_K|S)}{L(\bar{M}_{K+1}|S)}\right)}{\Gamma\left(\frac{d(K+1)-d(K)}{2}\right)} \tag{7}$$

where $d(K)$ are the *degrees of freedom* of model $\bar{M}_K$, Γ is the Euler Gamma function and γ is the lower incomplete gamma function.

The degrees of freedom of a Markov chain of order k over a state space $|V|$ are commonly given as $|V|^k(|V|-1)$ [1, 11, 29, 31]. This reflects that (i) the transition matrix of a Markov chain of order k has $|V|^{k+1}$ entries, and (ii) the rows in this matrix must sum to one. The latter reduces the free parameters by one for each of the $|V|^k$ rows, which yields the above expression. While this has been used to detect the Markov order in pathway data, this approach is not suitable for pathways that are constrained by a given (and known) network topology. It particularly neglects constraints due to the fact that not every sequence of vertices is a feasible path in a given network. As a simple example, for a graph consisting of two vertices A and B and a single directed edge (A, B), the vertex sequence $(A \to B \to A)$ is not a valid path of length two, even though the transition matrix of a second-order model contains a (zero) entry for the transition between (second-order) vertices (A, B) and (B, A). Hence, rather than calculating the degrees of freedom of a k-th order model based on the size of a transition matrix, we must only account for entries that correspond to paths in the underlying graph. The degrees of freedom of the k-th layer of a multi-order model thus depend on the number of

different paths of length k in a given graph G. For a binary adjacency matrix $\mathbf{A}$ of G, the entries $(\mathbf{A}^k)_{ij}$ in the k-th power of $\mathbf{A}$ count different paths of length k from i to j. Summing over the entries $(\mathbf{A}^k)_{ij}$ thus gives the total number of paths with length k. In the transition matrix of a k-th order model, we are free to set the entries corresponding to these paths, subject to the constraint that the matrix rows must sum to one. This reduces the degrees of freedom of a k-th order model by one for each *non-zero row* in the transition matrix. We thus get

$$\sum_{i,j}(\mathbf{A}^k)_{ij} - \sum_j \Theta\left(\sum_i (\mathbf{A}^k)_{ij} - 1\right) \tag{8}$$

where the sum $\sum\Theta(\cdot)$ over the Heaviside function Θ counts non-zero rows in $\mathbf{A}^k$. For a fully connected graph, the topology does not impose constraints on the possible paths of length k and in this case we recover the degrees of freedom of a standard Markov chain of order k.[1] Since a multi-order model combines higher-order models from $k = 0$ up to maximum order K, we sum the degrees of freedom of a zero-order model ($|V| - 1$) with Eq. 8 for $k \geq 1$:

$$d(K) = (|V| - 1) + \sum_{k=1}^{K}\left[\sum_{i,j}(\mathbf{A}^k)_{ij} - \sum_j \Theta\left(\sum_i (\mathbf{A}^k)_{ij} - 1\right)\right] \tag{9}$$

The difference between the degrees of freedom $d(K)$ of a multi-order model and standard higher-order Markov chains has important consequences for model selection: For sparse graphs (where a small fraction of possible edges exists) $d(K)$ calculated according to Eq. 9 increases considerably slower than the exponential increase expected for standard Markov chain models. This counters the curse of dimensionality, which has previously hindered the application of higher-order Markov models to pathway data [29].

In summary, we can detect the optimal maximum order K_{opt} of a multi-order graphical model by repeatedly calculating the p-value for consecutive pairs of (nested) models in the sequence $\bar{M}_1, \bar{M}_2, \ldots$. We then choose the maximum value $K_{opt} = K$ above which we reject the alternative model $\bar{M}_{K+1}$ in favor of $\bar{M}_K$, i.e. the largest K for which p is below a significance threshold ϵ. We note that, since the total number of likelihood ratio tests is K_{opt}, a small ϵ should be used to hinder false positives due to multiple hypothesis testing.

4.2 Experimental validation

We now validate our method using synthetically generated pathways. For this, we use a stochastic model generating a configurable number of variable-length paths, constrained by a random (directed) graph of variable size, based on a Markov chain with known order k. We omit the implementation details due to space constraints, however the full code of our model (along with all other code used in our work) is available in an online repository [24]. We then apply our method to these synthetically generated paths, showing that it (i) recovers the "correct" Markov order used to generate them, (ii) outperforms previously used Markov order detection techniques, and (iii) allows to infer an *optimal* higher-order graphical abstraction that can be used, e.g., to rank vertices.

Correctness and efficiency We compare our approach to two baseline Markov order detection techniques, which have previously been used to study pathways as *categorical sequences*. We specifically consider Markov order detection using (i) Aikake's Information Criterion (AIC) [32], and (ii) the Bayesian Information Criterion (BIC) [11, 29]. We apply both techniques to a single sequence of concatenated paths, where paths are separated by a special stop token (see, e.g., [19]). Fig. 2(a) compares the optimal maximum order K_{opt} inferred using our multi-order graphical models (MOG) to the order detected based on (BIC) and (AIC). Results are shown for different samples of N paths with known Markov order of four, generated in a toy random graph with 10 vertices and 30 directed edges. For moderately large samples AIC and BIC underfit the data, detecting the correct order only for $N > 50,000$ and $N > 350,000$ respectively, despite the small size of the graph. In contrast, our approach recovers the correct order for $N > 300$. We further recover the known result that BIC has a stronger tendency to underfit compared to AIC [11].

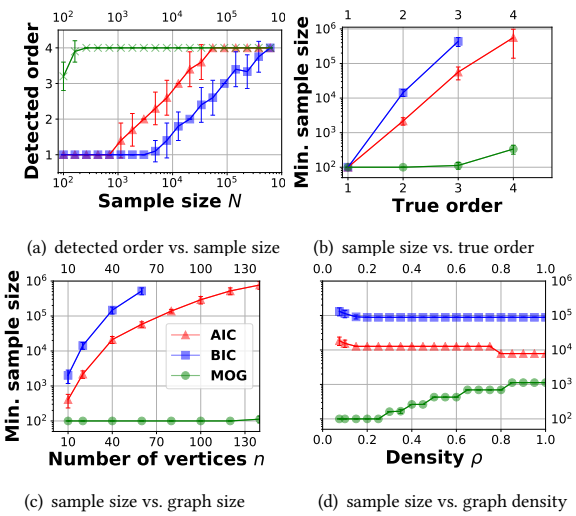


Figure 2: (a) shows detected Markov order (y-axis) for N synthetically generated paths (x-axis) and known Markov order four. (b-d) show the minimum sample size N (y-axis) needed to detect the correct Markov order for (b) paths in graphs with 20 vertices, 60 edges and with different Markov order (x-axis), (c) Markov order two, fixed edge density and varying graph size n (x-axis), and (d) Markov order two, graphs with 40 vertices and varying edge density ρ (x-axis). Results are averages of 20 experiments in random graphs, inferring the order based on Bayesian (BIC) and Aikake's (AIC) Information Criterion and Multi-order Graphical Models (MOG) proposed in this paper. Error bars indicate standard deviation.

We next study how the sample size N required to detect the correct order depends on (i) the Markov order, (ii) the number of vertices and (iii) the density of edges in the graph. Fig. 2(b) shows the results for different (true) Markov orders k used to generate paths in random graphs with 20 vertices and 60 directed edges. For AIC and BIC, N quickly grows for $k > 1$, while it remains small for our method. We further study how N depends on the size n (Fig. 2(c)) and density ρ (Fig. 2(d)) of the graph. As the number of vertices n in a sparse graph with $3n$ edges grows, the sample size needed by the BIC and AIC-based methods to detect the correct order two quickly exceeds $N = 10^6$. Our method yields the correct order also for small sample sizes (cf. Fig. 2(c)). We finally study how the minimally required sample size N depends on the density ρ of a graph with fixed size $n = 40$ (Fig. 2(d)). We define the density ρ as fraction of possible edges existing in a graph, i.e. $\rho = 0$ corresponds to an empty and $\rho = 1$ to a fully connected graph. As

[1] This follows from the fact that, for an $n \times n$ unit matrix $\mathbf{J} = (1)_{ij}$ of a fully connected graph, we have $\mathbf{J}^k = (n^{k-1})_{ij}$ and thus $\sum_{ij}\mathbf{J}^k_{ij} = n^2 \cdot n^{k-1} = |V|^k$. Since all $|V|^k$ rows in $\mathbf{J}^k$ are different from zero we recover $|V|^k(|V| - 1)$.

expected, the number of samples required by our method increases as the density, and thus the degrees of freedom of higher-order models, grow. For the BIC and the AIC we observe a mild decrease as the (real) degrees of freedoms in the fully connected graph approach those of a categorical sequence model. Interestingly, our method requires a smaller number of samples also for fully connected graphs, even though in this case the degrees of freedom of our model coincide with those used in the BIC and AIC-based methods. We attribute this to the fact that our method correctly accounts for multiple independent paths rather than aggregating them to a single sequence.

Ranking in Higher-Order Graphs We now show how our framework improves network-analytic methods, focusing on the ranking of vertices using PageRank [18]. We first recall that layer $k = 1$ of a multi-order model captures the topology of the graph and the (relative) frequencies of edges traversed by paths, while the layers $k > 1$ account for order correlations (of multiple lengths) that can break path transitivity. Hence, our framework can be viewed as a natural higher-order generalization of the common network abstraction of relational data, which not only captures the topology and frequency of links, but also order correlations in sequential data. From this perspective, the optimal maximum order K_{opt} allows to decide (i) if the (first-order) topology is sufficient to explain observed paths, or (ii) whether higher-order graphical models are needed. Moreover, we argue that K_{opt} is the *optimal* order of a higher-order graphical abstraction of pathway data.

To validate this claim, we use a set S of paths synthetically generated by the model above. The idea of our validation is to test whether the PageRank [18] calculated in a graphical model with order K_{opt} detected by our framework best captures the "ground truth" importance of vertices. For this, we recall that PageRank is a graph-based algorithm to calculate the stationary node visitation probabilities of random surfers in a (web) graph. In other words, it utilizes (i) the topology of the web graph, and (ii) a Markov chain model for random walks in the graph to estimate the (unknown) frequencies at which surfers visit web pages. Interpreting paths in our set S as trajectories of independent "surfers" in a graph, we can calculate the frequency at which a given vertex v is visited by these "surfers". With S_k denoting the multi-set of sub paths of length k, and considering that each vertex v is a zero-length sub path in S_0, we can thus calculate vertex "visitation frequencies" as

$$p_v = \frac{|\{v \in S_0\}|}{\sum_{p \in S} l_p + 1}. \tag{10}$$

The denominator simply counts all vertex traversals by summing up the number of vertices traversed by all paths $p \in S$. Interpreting p_v as the "ground truth" for the vertex visitation frequencies estimated by PageRank, we can subject the claim that our inference method yields an "optimal" graphical model of pathways to a numerical validation. For this, we generalize PageRank to a higher-order graph $G^{(k)}$ in a multi-order model. Let $\mathbf{A}^{(k)}$ be the binary adjacency matrix of $G^{(k)}$. We define $\mathbf{Q}^{(k)}$ as the matrix obtained by (i) dividing entries in $\mathbf{A}^{(k)}$ by row sums, and (ii) replacing zero rows by $1/n$, where n is the number of (higher-order) vertices in $G^{(k)}$. We then calculate a k-th order PageRank vector $x^{(k)}$ by solving the equation

$$x^{(k)} = x^{(k)} \left(\alpha \mathbf{Q}^{(k)} + (1 - \alpha) \mathbf{B} \right)$$

where $\mathbf{B}$ is an $n \times n$ matrix with entries $1/n$ and $\alpha = 0.85$ is a dampening factor. $x^{(k)}$ contains the PageRanks of k-th order vertices in $G^{(k)}$. Due to the De Bruijn graph construction (cf. Fig. 1), each k-th order vertex corresponds to a path $(v_0 \to \ldots \to v_{k-1})$ of length $k-1$. A projection to first-order vertices v can be defined as

$$\mathrm{PR}(k)[v] := \sum_{\substack{p \in S_{k-1} \\ v \sqsubseteq p}} \frac{1}{k} x_p^{(k)} \tag{11}$$

where $x_p^{(k)}$ is the PageRank of k-th order vertex p.[2] We can now test for which order k PR(k) best captures the ground truth visitation frequencies p_v calculated in a given synthetically generated set of pathways S. Fig. 3 shows the results for synthetically generated paths with different detected Markov orders (x-axis). Each of the five lines gives Kendall's rank correlation measure (y-axis) between a vertex ranking based on (i) "ground truth" visitation frequencies p_v calculated in actual pathways and (ii) the PageRank PR(k) for given order k. Naturally, the stochastic model underlying the PageRank calculation cannot perfectly reproduce the true frequencies at which vertices are "visited" by pathways. However, the results in Fig. 3 show that the PageRank in a k-th order graphical model reproduces ground truth visitation frequencies best if k corresponds to the optimal order K_{opt} detected by our framework. Thisconfirms that (i) a first-order PageRank yields suboptimal results for sequential data with order correlations, and (ii) the models learned by our approach are optimal to rank vertices.

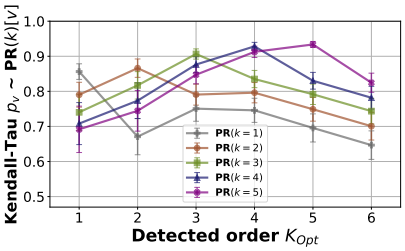


Figure 3: Kendall's rank correlation between k-th order PageRank $PR(k)$ and ground truth vertex visitation frequencies p_v (y-axis) in paths with different detected orders K_{opt} (x-axis). Results are averages of 100 runs, fitting a multi-order model to $N = 20000$ synthetically generated paths of length $L = 10$ in random graphs with 100 vertices and 350 edges. Error bars indicate standard deviation.

5 APPLICATIONS

Having validated our method in synthetic examples, we now apply it to eight real data sets from different scenarios: We start with data that provide pathway statistics, namely (i) passenger itineraries in transportation networks, (ii) click streams of users on the Web, and (iii) career paths of scientists. We then show how our method can be used to analyze time-stamped interactions, commonly studied as *temporal* or *dynamic networks*. Key characteristics and sources of the data sets are shown in Table 1. All are freely available for research and details on how they have been collected are introduced along the way. Results in this section have been obtained using pathpy, an OpenSource `python` implementation of our framework [25]. The full code of our analysis is available online [24].

5.1 Pathway Data

We study five pathway data sets: (AIR) captures 280k passenger itineraries along flight routes between US airports in 2001 [27, 33], (TUBE) contains 4.2 million passenger trips in the London metro [7, 27], (CAREER)

[2] Since $x^{(k)}$ is a stochastic vector, Eq. 11 ensures that entries of PR(k) sum to one.

Pathway Data	**Vertices ($\|V\|$)**	**Edges ($\|E\|$)**	**Paths (N)**	**[Min, Max] l_i**	**K_{opt} (p-value)**
Scientist career paths (CAREER) [30]	1,932 (institutes)	6,474	33,576	[0, 12]	1 ($p \approx 0$)
Wikispeedia click paths (WIKI) [35]	100 (Wikipedia pages)	1,790	39,846	[0, 21]	2 ($p \approx 0$)
US airflight itineraries (AIR) [33]	175 (US airports)	1,598	286,810	[1, 13]	2 ($p \approx 0$)
MSNBC clickstreams (MSNBC) [3]	17 (page categories)	289	989,818	[0, 99]	3 ($p \approx 0$)
London Tube itineraries (TUBE) [7]	276 (metro stations)	663	4,295,731	[1, 35]	6 ($p \approx 0$)
Temporal Network Data	**Vertices ($\|V\|$)**	**Edges ($\|E\|$)**	**Paths (N)**	**δ/[Min, Max] l_i**	**K_{opt} (p-value)**
Company E-Mails (EMAIL) [17]	167 (employees)	5,784	80,504	30/[1, 13]	1 ($p \approx 0$)
Workplace Contacts (WORK) [10]	92 (office workers)	755	10,939	180/[1, 4]	2 ($p \approx 0$)
Hospital Contacts (HOSP) [34]	75 (healthcare workers)	1,139	353,449	300/[1, 9]	3 ($p \approx 0$)

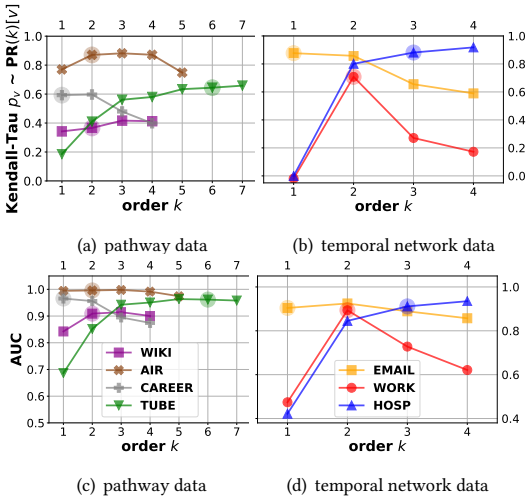
Table 1: Summary statistics and detected maximum order K_{opt} of multi-order graphical model for real-world data sets.

contains sequences of affiliations in the career of more than 30k scientists publishing in journals of the American Physical Society [30], and (WIKI) provides more than 76k click paths of users playing the Wikispeedia navigation game [35]. For (WIKI) the small number of observed paths compared to size and density of the underlying Wikipedia article graph renders a detection of higher Markov orders impossible.[3] To overcome this problem, we limit our analysis to click paths that traverse the 100 most frequently visited articles. We finally consider (MSNBC), a data set with close to one million click streams of visitors of the MSNBC portal [3].

For (AIR), (TUBE), and (WIKI) observed pathways are, by definition, constrained to an underlying network of available flight routes, London metro lines, and Wikipedia article links used in the Wikispeedia game respectively. For (CAREER), the situation is more difficult: On the one hand, researchers can, in principle, move between any pair of affiliations. On the other hand, geographic locations, research disciplines, and hiring strategies of affiliations render some of these theoretically possible affiliation changes unlikely (or even impossible). For the following analysis we thus take a simple approach, assuming that affiliation changes are constrained to those that have been observed at least once. Finally, (MSNBC) contains user click streams at the level of page *categories*. Different from (WIKI), these click streams are *not* constrained to paths in a given (article or web) graph. For (MSNBC) we thus assume a fully connected graph topology, for which the degrees of freedom in our model coincide with those commonly used in standard Markov order detection techniques (cf. section 4.1). We still include (MSNBC) in our analysis to confirm that, for such *unconstrained* path, our method recovers the results reported in [29].

For each data set, we first learn a multi-order graphical model, inferring the maximum order K_{opt} as described in section 4 (using $\epsilon = 0.001$). Notably, BIC and AIC-based order detection yield order one for all data sets, except for (MSNBC) where both recover order three thanks to a small number of categories and large sample size. Table 1 shows that, in contrast, our method yields $K_{opt} > 1$ for all data sets except for (CAREER). This indicates that a first-order network model is not justified for four of the five data sets. We validate this using the approach introduced in section 4.2, i.e. we use pathways to calculate ground truth vertex visitation frequencies p_v and check for which order k a k-th order PageRank best recovers this ground truth.[4] Fig. 4(a) reports Kendall's rank correlation coefficient (τ) between a ranking obtained from (i) ground truth visitation frequencies p_v and (ii) PageRank $PR(k)$ computed for different k as described in 4.2. While the extent to which PageRank reproduces this ground truth naturally varies, our results confirm that K_{opt} inferred by our method is indeed the "optimal" order of a graphical model: For (CAREER), where our method yields $K_{opt} = 1$, we observe a maximum $\tau \approx 0.59$ for $k = 1$, while τ drops for $k > 2$. For (AIR) and (TUBE) τ increases for $k > 1$, saturating at the detected orders $K_{opt} = 2$ and $K_{opt} = 6$ respectively. We highlight that a first-order model of (TUBE) yields misleading results, which raises interesting questions about network-based studies of transportation systems. Interestingly, increasing k beyond K_{opt} does not necessarily decrease τ. For (TUBE) and (WIKI) we even observe slight increases of τ for $k > K_{opt}$. However, since our method accounts for model complexity it correctly determines the order K_{opt} beyond which additional layers are not justified by the (small) increase in "explanatory power".

To corroborate this interpretation, we study the predictive power of higher-order graphical models. Here, we want to predict most frequently visited vertices, i.e. vertices v for which p_v is largest. Our prediction is based on the top-ranked vertices according to PageRank, calculated in graphical models with different orders k. For each k this yields a predictor for which we calculate the Area under the Curve (AUC) shown in Fig. 4(c). For (CAREER), where we infer $K_{opt} = 1$, higher-order models do not yield better predictions than a first-order model. For (TUBE), the performance of a first-order model is low ($AUC(1) \approx 0.69$), while we find $AUC(K_{opt} = 6) \approx 0.96$. For (TUBE) and (WIKI) we find that, despite τ slightly increasing for $k > K_{opt}$, such larger k do not translate to better predictions. For (WIKI) AUC increases considerably in a second-order model, even though τ shows only a small increase. This confirms that the predictive quality of PageRank is optimal for graphical models with order K_{opt}.



(a) pathway data (b) temporal network data

(c) pathway data (d) temporal network data

Figure 4: (a-b) show Kendall's rank correlation between visitation frequencies p_v and PageRank (y-axis) calculated in k-th order model for different k (x-axis) in pathways (a) and temporal networks (b). (c-d) show Area Under Curve (AUC) for prediction of 15% most frequently visited vertices based on PageRank computed for different k. Values k that correspond to the detected order K_{opt} are highlighted (cf. Table 1).

[3] Note that the small sample size in (WIKI) also poses challenges for variable Markov order modeling techniques as well as for AIC/BIC-based Markov order detection.

[4] Since it only provides data on 17 page categories connected via a (trivial) fully connected topology, we omit this analysis for (MSNBC).

5.2 Temporal Network Data

Apart from settings where we have access to pathway data, we finally discuss how our framework can be applied to *time-stamped data* on *temporal or dynamic networks*. I.e., we consider triplet data of the form $(v, w; t)$, which capture that two vertices v and w were connected at (discrete) time t. Despite their growing importance, e.g., in social network analysis, analyzing such data is still a challenge [9]. A number of works have shown that standard network-analytic and algebraic methods yield wrong results, e.g., about dynamical processes, centralities or cluster structures in temporal networks [15, 19–22, 27, 38]. These limitations have been attributed to temporal correlations in the sequence of edges and their effect on so-called *time-respecting paths* [12]. We consider a sequence $(v_0, v_1; t_1), (v_1, v_2; t_2), \ldots, (v_{l-1}, v_l; t_l)$ of time-stamped edges a *time-respecting path* $(v_0 \rightarrow \ldots \rightarrow v_l)$ iff the ordering of edges respects causality, i.e. $t_1 < \ldots < t_l$. Importantly, this implies that the *ordering* of time-stamped edges can invalidate the transitivity of paths implicitly assumed by time-aggregated analyses: Specifically, two time-stamped edges $(A, B; t)$ and $(B, C; t')$ give rise to a transitive path $(A \rightarrow B \rightarrow C)$ only if $(A, B; t)$ occurs *before* $(B, C; t')$. Hence, correlations in the ordering of edges can break transitivity and thus invalidate network-analytic methods [15, 20, 27].

We now show that our framework (i) detects these correlations and (ii) infers a multi-order graphical model that captures both temporal and topological characteristics of temporal networks. For this, we follow the common approach and consider – in addition to their ordering - the actual *timing* of time-stamped edges in the definition of time-respecting paths [9]. We particularly require that edge sequences contributing to time-respecting paths are consistent with a *maximum time difference* δ between consecutive edges, i.e. $0 < t_{i+1} - t_i \leq \delta$ $(i = 0, \ldots, l)$. This is important since we are typically interested in paths that mediate processes evolving at time scales much shorter than the observation period [9]. With this definition of a time-respecting path at hand, we apply the following procedure: We first use time-stamped edges to extract time-respecting paths for a given δ, obtaining a multi-set of (time-respecting) paths S. We then use the method discussed in section 4 to infer a multi-order model, where (i) layers $k = 0$ and $k = 1$ model "activities" of vertices as well as the topology and frequency of time-stamped edges and (ii) layers $k > 1$ capture correlations in the ordering of edges that influence longer (time-respecting) paths. $K_{opt} > 1$ indicates that these correlations invalidate a (first-order) network abstraction. In this case, K_{opt} further provides us with the optimal order of a (higher-order) graphical representation.

We apply this to three temporal network data sets, summarized in Table 1: (EMAIL) captures time-stamped E-Mail exchanges between 167 company employees [17], (HOSP) contains time-stamped contacts between 75 healthcare workers in a hospital [34], and (WORK) captures time-stamped contacts between 92 office workers [8]. (HOSP) and (WORK) were recorded using badges sensing face-to-face encounters at high temporal resolution [8, 34]. For each data set we first extract time-respecting paths for a given maximum time difference δ. The optimal choice of δ is a difficult research problem by itself. Here we use a simple approach, choosing δ based on the inter-event time distribution (which captures "inherent" time scales of the data, cf. Table 1). We then infer the optimal maximum order K_{opt} of a multi-order model. Table 1 shows that a first-order model is justified for (EMAIL), while (HOSP) and (WORK) exhibit temporal correlations that warrant higher-order models. We subject the intuition that correlations in the ordering of edges necessitate higher-order models to a simple sanity check: We randomly shuffle time stamps of edges to destroy temporal correlations, extract time-respecting paths for the shuffled data, and again infer the optimal maximum order of a multi-order model. We get $K_{opt} = 1$ for all shuffled data sets, confirming that first-order graphical abstractions of temporal networks are justified only if temporal correlations in the sequence of time-stamped edges are absent.

Our results indicate that first-order network models of (HOSP) and (WORK) likely yield wrong results, while they seem justified in (EMAIL). We again validate this by checking the correlation between (i) ground truth vertex visitation frequencies by time-respecting paths and (ii) the PageRank $PR(k)$ calculated for different orders k. Like above, we study the AUC of higher-order PageRanks for different orders k. Fig. 4(b) shows that for higher-order models with $k > 1$ the rank correlation does not increase for (EMAIL), while it strongly increases for (HOSP) and (WORK). For the latter two, first-order PageRank is uncorrelated with the ground truth, while graphical models with order K_{opt} yield $\tau \approx 0.71$ and $\tau \approx 0.67$ respectively. For (HOSP) and (WORK), Fig. 4(d) shows a strong increase of AUC for $PR(K_{opt})$ to values of 0.91 and 0.89 respectively. For (EMAIL) we observe no increase. We attribute this to strong temporal correlations in (HOSP) and (WORK), which affect time-respecting paths and render first-order network abstractions useless. This confirms (i) that the optimal order inferred by our method is meaningful and (ii) that it allows to decide when a first-order representation of time series data is justified.

6 CONCLUSION

Graph- and network-analytic methods are widely applied to data that capture relations between elements. While researchers in data science raised concerns about their application to data with complex characteristics, we lack principled methods to decide when network abstractions are justified and when not. Addressing this issue, we propose a solution for data on pathways and temporal networks. Going beyond previous works, we generalize common network abstractions to multi-order graphical models. We advance the state-of-the-art in sequential data mining by proposing a model selection technique that accounts for the characteristics of data carrying multiple observations of paths in a graph. A comparison to previously used methods shows that it considerably improves the inference of *optimal* graphical models that balance model complexity and explanatory power. These models can be seen as optimal graphical "summarizations" of sequential data, which can be used to improve network analysis and modeling techniques. We demonstrate the relevance of our method in real data on click streams, career paths, and transportation networks. We highlight implications for the study of temporal networks, which are often analyzed using time-aggregated or time-slice graphs. We show that temporal correlations invalidate such analyses and demonstrate that our method can be used to learn optimal graph models that capture temporal and topological characteristics of time series data.

In conclusion, our work highlights fallacies of network abstractions of sequential data. Principled model selection is a crucial first task that must precede any application of network-analytic methods. The proposed framework is a step in this direction. It points out relations between network analysis and sequential pattern mining that call for further research. To facilitate its application and to ensure the reproducibility of our results, an OpenSource python implementation of our framework is available [25].

REFERENCES

[1] Theodore W Anderson and Leo A Goodman. 1957. Statistical inference about Markov chains. *The Annals of Mathematical Statistics* (1957), 89–110.

[2] Carter T Butts. 2009. Revisiting the foundations of network analysis. *science* 325, 5939 (2009), 414–416.

[3] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. 2000. Visualization of Navigation Patterns on a Web Site Using Model-based Clustering. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)*. 280–284.

[4] Flavio Chierichetti, Ravi Kumar, Prabhakar Raghavan, and Tamas Sarlos. 2012. Are Web Users Really Markovian?. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 609–618.

[5] N. G. de Bruijn. 1946. A Combinatorial Problem. *Koninklijke Nederlandse Akademie v. Wetenschappen* 49 (1946), 758–764.

[6] Alceu Ferraz Costa, Yuto Yamaguchi, Agma Juci Machado Traina, Caetano Traina, Jr., and Christos Faloutsos. 2015. RSC: Mining and Modeling Temporal Activity in Social Media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 269–278.

[7] Transport for London. 2014. Rolling Origin and Destination Survey (RODS) database. (2014). http://www.tfl.gov.uk/info-for/open-data-users/our-feeds

[8] Mathieu Génois, Christian L Vestergaard, Julie Fournet, André Panisson, Isabelle Bonmarin, and Alain Barrat. 2015. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science* 3, 03 (2015), 326–347.

[9] Petter Holme. 2015. Modern temporal network theory: a colloquium. *The European Physical Journal B* 88, 9 (2015), 234.

[10] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. 2011. What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theo. Biol.* 271, 1 (2011), 166–180.

[11] Richard W Katz. 1981. On some criteria for estimating the order of a Markov chain. *Technometrics* 23, 3 (1981), 243–249.

[12] David Kempe, Jon Kleinberg, and Amit Kumar. 2000. Connectivity and inference problems for temporal networks. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*. ACM, 504–513.

[13] Taehwan Kim, Yisong Yue, Sarah Taylor, and Iain Matthews. 2015. A Decision Tree Framework for Spatiotemporal Sequence Prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 577–586.

[14] Ravi Kumar, Maithra Raghu, Tamás Sarlós, and Andrew Tomkins. 2017. Linear Additive Markov Processes. In *Proceedings of the 26th International Conference on World Wide Web*. 411–419.

[15] Hartmut H. K. Lentz, Thomas Selhorst, and Igor M. Sokolov. 2013. Unfolding Accessibility Provides a Macroscopic Approach to Temporal Networks. *Phys. Rev. Lett.* 110 (Mar 2013), 118701. Issue 11.

[16] Chuanren Liu, Kai Zhang, Hui Xiong, Geoff Jiang, and Qiang Yang. 2014. Temporal Skeletonization on Sequential Data: Patterns, Categorization, and Visualization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. 1336–1345.

[17] R. Michalski, S. Palus, and P. Kazienko. 2011. Matching Organizational Structure and Social Network Extracted from Email Communication. In *Lecture Notes in Business Information Processing*. Vol. 87. Springer Berlin Heidelberg, 197–206.

[18] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford.

[19] Tiago P Peixoto and Martin Rosvall. 2015. Modeling sequences and temporal networks with dynamic community structures. *ArXiv* (2015). arXiv:1509.04740

[20] René Pfitzner, Ingo Scholtes, Antonios Garas, Claudio J Tessone, and Frank Schweitzer. 2013. Betweenness Preference: Quantifying Correlations in the Topological Dynamics of Temporal Networks. *Phys. Rev. Lett.* 110 (May 2013).

[21] Martin Rosvall, Alcides V Esquivel, Andrea Lancichinetti, Jevin D West, and Renaud Lambiotte. 2014. Memory in network flows and its effects on spreading dynamics and community detection. *Nat. Comm.* 5 (Aug 2014).

[22] V. Salnikov, M. T. Schaub, and R. Lambiotte. 2016. Using higher-order Markov models to reveal flow-based communities in networks. *Sci. Rep.* 6, Article 23194 (2016), 23194 pages.

[23] Ramesh R. Sarukkai. 2000. Link Prediction and Path Analysis Using Markov Chains. *Comput. Netw.* 33, 1-6 (June 2000), 377–386.

[24] Ingo Scholtes. 2017. Multi-Order Graphical Modeling of Pathways and Temporal Networks: Supplementary Code. (2017). https://doi.org/10.5281/zenodo.293010

[25] Ingo Scholtes. 2017. python package pathpy. https://github.com/IngoScholtes/pathpy. (2017).

[26] Ingo Scholtes, Nicolas Wider, and Antonios Garas. 2016. Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities. *The European Physical Journal B* 89, 3 (2016), 61.

[27] Ingo Scholtes, Nicolas Wider, René Pfitzner, Antonios Garas, Claudio J Tessone, and Frank Schweitzer. 2014. Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nat. Comm.* 5 (Sept 2014), 5024.

[28] Gideon Schwarz. 1978. Estimating the dimension of a model. *The annals of statistics* 6, 2 (1978), 461–464.

[29] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. 2014. Detecting Memory and Structure in Human Navigation Patterns Using Markov Chain Models of Varying Order. *PLoS ONE* 9, 7 (07 2014), 1–21.

[30] American Physical Society. 2016. (2016). https://journals.aps.org/datasets

[31] Christopher C Strelioff, James P Crutchfield, and Alfred W Hübler. 2007. Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Phys. Rev. E* 76 (Jul 2007), 011106. Issue 1.

[32] Howell Tong. 1975. Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability* 12, 03 (1975), 488–497.

[33] RITA TransStat. 2014. Origin and Destination Survey database. (2014). http://www.transtats.bts.gov/Tables.asp?DB_ID=125

[34] P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Regis, B.-A. Kim, B. Comte, and N. Voirin. 2013. Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. *PLoS ONE* 8 (2013).

[35] Robert West and Jure Leskovec. 2012. Human Wayfinding in Information Networks. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 619–628.

[36] Samuel S Wilks. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 9, 1 (1938).

[37] T. Wu and D. Gleich. 2017. Retrospective Higher-Order Markov Processes for User Trails. *ArXiv* (2017). arXiv:1704.05982

[38] Jian Xu, Thanuka L Wickramarathne, and Nitesh V Chawla. 2016. Representing higher-order dependencies in networks. *Science advances* 2, 5 (2016), e1600028.

[39] Y Yang, D Yan, H Wu, J Cheng, S Zhou, and J CS Lui. 2016. Diversified Temporal Subgraph Pattern Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 1965–1974.

[40] Hongyang Zhang, Peter Lofgren, and Ashish Goel. 2016. Approximate Personalized PageRank on Dynamic Graphs. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 10.

[41] Yan Zhang, Antonios Garas, and Ingo Scholtes. 2017. Controllability of temporal networks: An analysis using higher-order networks. *ArXiv* (2017). arXiv:physics.soc-ph/1701.06331

[42] Jacob Ziv and Abraham Lempel. 1977. A universal algorithm for sequential data compression. *IEEE Transactions on information theory* 23, 3 (1977), 337–343.

[43] Katharina A Zweig. 2011. Good versus optimal: Why network analytic methods need more systematic evaluation. *Central Europ. J. Computer Science* 1, 1 (2011), 137–153.