Analysis and Visualisation of Time Series Data on Networks with `Pathpy`

Jürgen Hackl
hackl@liverpool.ac.uk
University of Liverpool
Liverpool, United Kingdom

Ingo Scholtes
scholtes@uni-wuppertal.de
University of Wuppertal
Wuppertal, Germany

Luka V. Petrović
petrovic@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Vincenzo Perri
perri@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Luca Verginer
lverginer@ethz.ch
ETH Zurich
Zurich, Switzerland

Christoph Gote
cgote@ethz.ch
ETH Zurich
Zurich, Switzerland

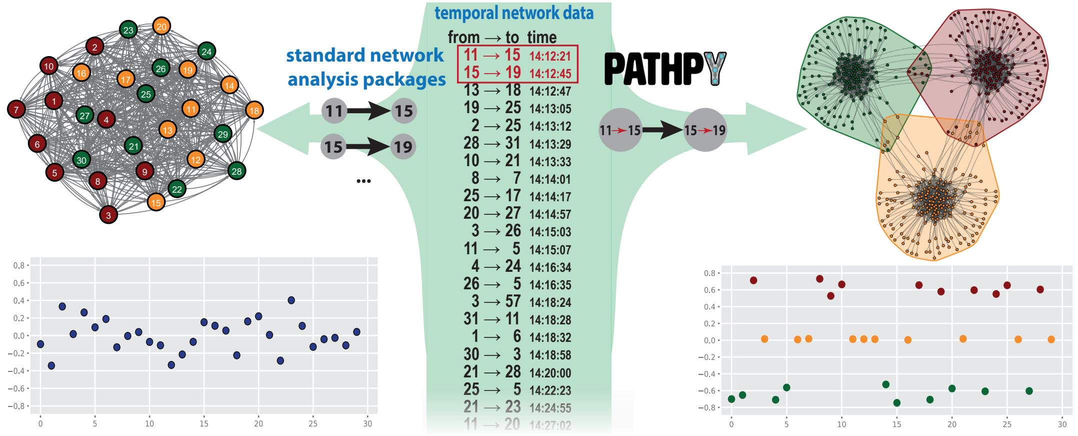

Figure 1: Exemplary application of `pathpy` to detect clusters in time-stamped network data (center). Common network analysis packages are based on time-aggregated graph representations (top left) that decompose sequences of interactions that constitute a causal paths into independent dyadic links. The Fiedler vector of such a graph (bottom left) does not allow to detect temporal-topological clusters (colored nodes) introduced by the temporal ordering of links. In contrast, `pathpy` uses statistical learning techniques to infer optimal higher-order graph models that capture non-dyadic dependencies (top right) generated by the statistics of causal paths. Such models allow, e.g., to detect temporal-topological clusters via a generalization of spectral clustering to higher-order Laplacians (bottom right).

ABSTRACT

The Open Source software package `pathpy`, available at https://www.pathpy.net, implements statistical techniques to learn optimal graphical models for the causal topology generated by paths in time-series data. Operationalizing Occam's razor, these models balance model complexity with explanatory power for empirically observed paths in relational time series. Standard network analysis is justified if the inferred optimal model is a first-order network model. Optimal models with orders larger than one indicate higher-order dependencies and can be used to improve the analysis of dynamical processes, node centralities and clusters.

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452052

CCS CONCEPTS

• **Computing methodologies → Machine learning approaches**; **Network science**; • **Networks → Network algorithms**; • **Software and its engineering → Software notations and tools**.

KEYWORDS

higher-order graph models, temporal network, graph mining, network analysis, network visualization, causal paths, software

ACM Reference Format:
Jürgen Hackl, Ingo Scholtes, Luka V. Petrović, Vincenzo Perri, Luca Verginer, and Christoph Gote. 2021. Analysis and Visualisation of Time Series Data on Networks with `Pathpy`. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3442442.3452052

1 INTRODUCTION

Network-based data mining techniques such as graph mining, (social) network analysis, link prediction and graph clustering are a cornerstone for data science and machine learning in Web data. They help to detect patterns in large data sets that capture *dyadic relations* between pairs of documents, websites, users, or products and have improved our understanding of complex networks across disciplines. While the potential of analysing *graph* or *network models* of relational data is undisputed, we increasingly have access to *temporal* data that not only tell us *who* is related to *whom* but also *when* and *in which order* relations occur. Consider, e.g., data on user clickstreams in the Web, time-stamped social networks, or sequences of word co-occurrences in documents. Such temporal and sequential data pose a fundamental challenge for state-of-the-art graph mining and network analysis techniques. Aggregating data within certain time slices, most state-of-the-art techniques discard information on the microscopic timing and ordering of links, which is, however, the foundation of so-called *time-respecting* or *causal paths* [1]. That is, we lose information on *who* can influence *whom* indirectly. For a sequence of two links $A \longrightarrow B$ and $B \longrightarrow C$, a node A can only influence C via a (transitive) *causal path* via B if $A \longrightarrow B$ occurs *before* $B \longrightarrow C$. A causal path from A to C does not exist if the order of links is reversed. Recent works show that the timing and ordering of interactions in real systems can introduce higher-order, non-dyadic dependencies not captured by state-of-the-art graph models [2]. We still lack software packages that enable us to analyse this important dimension of complexity in relational time series. Addressing this gap, we present pathpy, a python package that provides data analysis and machine learning methods for temporal networks. It is suitable for researchers and practitioners who (i) wish to adopt a graph perspective to reveal higher-order dependencies between elements of complex systems, (ii) extract higher-order features for machine learning tasks on relational time series, or (iii) need methods that go beyond the dyadic perspective of existing graph mining and network analysis tools.

2 KEY PATHPY FEATURES

Pathpy is an actively developed library with a growing number of contributors and users as well as scalable data analytics and machine learning techniques for temporal networks. It is specifically tailored to time-stamped and sequential data that capture multiple short paths observed in a graph. Examples for such data in the context of the World Wide Web include dynamic social interactions, user click streams, citation graphs, or traces of information propagation in social media. Unifying the analysis of such temporal data, pathpy provides efficient methods to calculate statistics of causal paths. In the remainder of this section, we highlight selected pathpy features.

Higher-Order Graph Models. The foundation of pathpy are so-called higher- and muti-order graph models, a framework that generalises standard graph representations of relational data to k-dimensional De Bruijn graph models for causal paths in temporal data [5, 6]. Standard time-aggregated graph models of temporal networks can be viewed as one-dimensional De Bruijn graphs, where link weights capture the frequencies of links (which are causal paths of length $k = 1$) between nodes (which can be viewed as paths of length $k - 1 = 0$). Generalizing this idea, weighted links in a k-th order graph model capture frequencies of causal paths of length k between nodes representing paths of length $k - 1$. Fig. 1 shows an example for a second-order model, where the indicated link represents a path of length $k = 2$ consisting of two consecutive interactions $11 \longrightarrow 15$ and $15 \longrightarrow 19$. Such higher-order graph models have proven to be a poweful approach to understand the *causal topology* of complex systems [2]. They can be used to generalize network analysis and graph mining to temporal data, and help us to address limitations of social network analysis techniques.

Optimal Order Detection. A fundamental question in the modelling of data via higher-order graph models is which higher order k is needed to analyse a given time series. For some data, standard graph representations (i.e. first-order model with $k = 1$) are sufficient while others require models with larger order $k > 1$. To answer this crucial question in the modelling and analysis of temporal Web data, pathpy implements model selection and statistical learning techniques that allow to (i) decide when standard graph representations of time series data are justified, and (ii) determine the *optimal* order of higher-order models for data that cannot be modelled as (first-order) graphs [5].

Optimal Higher-Order Graph Analytics. Apart from techniques to learn optimal higher-order graph models, pathpy implements generalizations of key graph analytic methods to those higher-order models. Examples include algorithms to rank nodes based on different notions of centralities such as betweenness, closeness, eigenvector, or PageRank centrality defined in higher-order models [5, 6], methods to compute stationary states, visitation probabilities, and convergence times of higher-order random walk models that are the foundation for time-aware node embedding or clustering algorithms [4], as well as spectral analysis techniques building on higher-order generalizations of Laplacian matrices [7].

Interactive Visualisations. Pathpy is fully integrated with jupyter, providing rich, fully-customizable and interactive visualisations of static, temporal, and higher-order networks. Building on a recently developed layout algorithm [3], time-aware static visualizations of temporal graphs can be generated which highlight patterns in the underlying temporal data. Visualisations can be exported to HTML5 files that can easily be shared and published on the Web, or converted to tex/tikz code that faciliates customizable figures suitable for scientific publications.

3 CONCLUSION

In summary, pathpy provides easy access to an array of network analysis and machine learning techniques for temporal data on networks. Building on robust, scalable, and easy-to-use data structures that are coherent with python's data science stack, pathpy is a compelling choice for data science tasks in temporal Web data such as time-stamped social networks or clickstreams. More information on pathpy is available at https://www.pathpy.net.

ACKNOWLEDGMENTS

We thank all contributors and users of pathpy for their feedback, bug reports, feature requests, and code contributions.

REFERENCES

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

[2] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. 2019. From networks to optimal higher-order models of complex systems. *Nature physics* 15, 4 (2019), 313–320.

[3] Vincenzo Perri and Ingo Scholtes. 2020. HOTVis: Higher-Order Time-Aware Visualisation of Dynamic Graphs. In *Proceedings of the 28th International Symposium on Graph Drawing and Network Visualization (Graph Drawing 2020), Vancouver, BC, Canada, in press.* arXiv:1908.05976 [cs.SI]

[4] 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. *Nature communications* 5 (2014), ncomms5630.

[5] Ingo Scholtes. 2017. When is a Network a Network?: Multi-Order Graphical Model Selection in Pathways and Temporal Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. ACM, New York, NY, USA, 1037–1046. https://doi.org/10.1145/3097983.3098145

[6] 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 (02 Mar 2016), 61. https://doi.org/10.1140/epjb/e2016-60663-0

[7] Ingo Scholtes, Nicolas Wider, Rene Pfitzner, Antonios Garas, Claudio Juan Tessone, and Frank Schweitzer. 2014. Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature Communications* 5 (September 2014), 5024. https://doi.org/10.1038/ncomms6024 arXiv:1307.4030