Higher-Order Network Models for Temporal Network Data

Jürgen Hackl

Civil Engineering and Industrial Design
University of Liverpool, UK

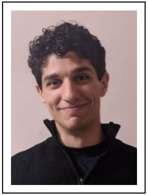
Data Analytics Group
University of Zurich, CH

hackl@liverpool.ac.uk

Introductory Lecture
Tutorial @ NetSciX 2020 Tokyo

2020/01/23

Data Analytics Group (DAG)



Vincenzo Perri

Luka Petrovic

Prof. Dr. Jürgen Hackl

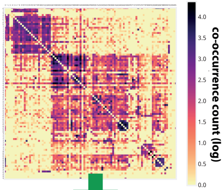
Prof. Dr. Ingo Scholtes

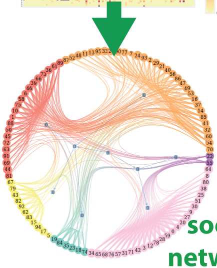
Data Analytics team at Department of Informatics at UZH

Data analytics in complex systems

- **Computational Social Science**
  1. resilience of Open Source communities
     Cooperative and Human Aspects of Software Engineering, May 2013
  2. social factors in citation networks
     EPJ Data Science, March 2014
  3. quantifying social influence in peer review
     Scientometrics, March 2017
- **Team-based Software Development**
  4. automated classification of bug reports
     Int. Conf. on Software Engineering (ICSE), May 2013
  5. team structures and team performance
     Empirical Software Engineering, April 2016
  6. mining networks from git repositiories
     Mining Software Repositories, March 2019
- **Data Analytics for Complex Relational Data**
  7. network inference from noisy sensor data
     SocInfo, August 2017
  8. network analytics for time series data
     SIGKDD, August 2017
  9. higher-order models for paths
     Nature Physics, March 2019

noisy observational data





social network?

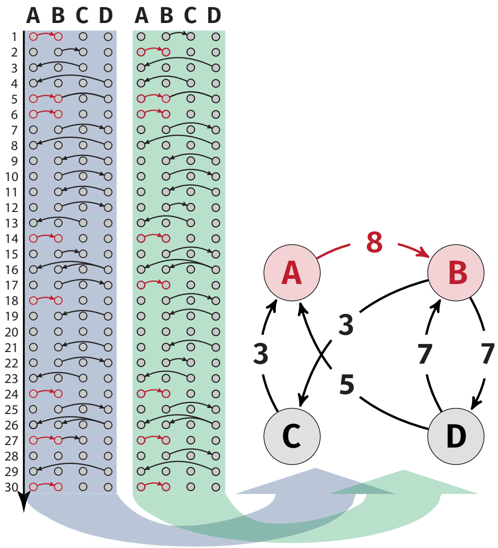
Network models?

from → to	when
10 → 15	14:12:21
43 → 38	14:12:47
22 → 57	14:12:49
19 → 35	14:13:05
2 → 25	14:13:12
48 → 31	14:13:29
30 → 21	14:13:33
8 → 7	14:14:01
55 → 17	14:14:17
20 → 27	14:14:57
3 → 40	14:15:03
51 → 5	14:15:07
4 → 24	14:16:34
56 → 5	14:16:35
3 → 57	14:18:24
31 → 11	14:18:28
1 → 6	14:18:32
40 → 3	14:18:58
31 → 48	14:20:00
25 → 5	14:22:23
21 → 43	14:24:55
41 → 30	14:27:02
11 → 31	14:29:46
5 → 56	14:30:01
35 → 2	14:31:04
15 → 10	14:31:29

problem

- temporal information can **invalidate network models**
- **limits network analysis** of time-stamped social interactions

Why is time important?



arrow of time

- chronological ordering of links determines **causal paths**
- network models **discard information on causal topology**

⟶ R Pfitzner, I Scholtes et al., Phys Rev Lett 2013

⟶ HK Lentz et al., Phys Rev Lett 2013

Ordering matters!

- graph analytic & algebraic methods: **transitive & Markovian paths**, i.e.

$$P(\overrightarrow{ABC}) = P(\overrightarrow{AB}) \cdot P(\overrightarrow{BC})$$

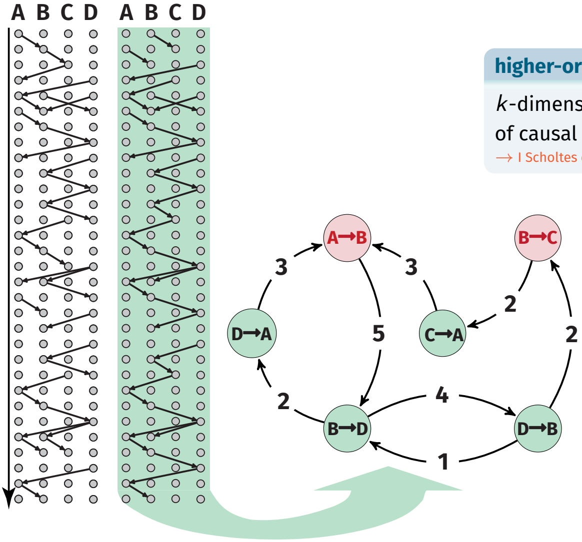
- temporal correlations lead to **non-Markovian paths**
- real time series data exhibit **higher-order dependencies**

misleading results about ...

- path structures
- random walks
- cluster structures
- node centralities

$\longrightarrow$ R Pfitzner, I Scholtes et al., Phys Rev Lett 2013
$\longrightarrow$ I Scholtes et al., Nature Comm. 2014
$\longrightarrow$ M Rosvall et al., Nature Comm. 2014
$\longrightarrow$ I Scholtes et al., SIGKDD 2017

From graphs to higher-order models



higher-order network model

k-dimensional ***De Bruijn* graph model** of causal paths

⟶ I Scholtes et al., Nature Comm. 2014

Time-aware node ranking

centrality measures

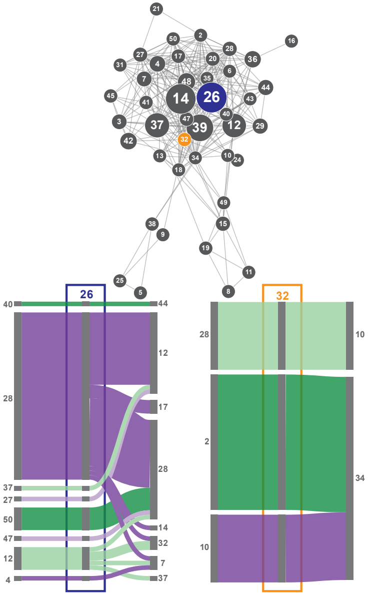
- find “influential” individuals
  - **spectral centralities**
  - **path-based centralities**
- <u>but:</u> focus on **topological** importance

temporal-topological centrality

- ordering of links changes **temporal influence** of nodes → H Habiba et al., 2007
- **higher-order centralities** capture influential nodes in temporal networks
  → I Scholtes, N Wider, A Garas, EPJ B 2016

in practice

time-stamped interactions between software developers in OpenSource project Eclipse



shortest causal paths
in **time series data**

Time-aware graph clustering

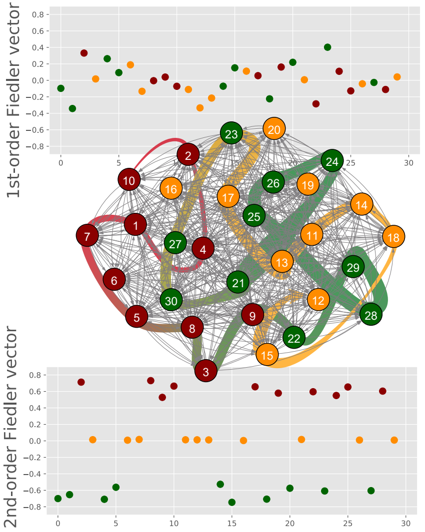
graph clustering algorithms

- fnd clusters of “well-connected” nodes
  - **modularity optimisation**
  - **clique percolation**
  - **flow compression**
  - **spectral clustering**
- <u>but:</u> focus on **topological** communities

temporal-topological clusters

- ordering of links can introduce **clusters in causal topology**
- spectral clustering in **higher-order Laplacians** → I Scholtes et al., Nature Comm. 2014
- **higher-order flow compression**

  → D Edler, L Bohlin, M Rosvall, Algorithms, 2017



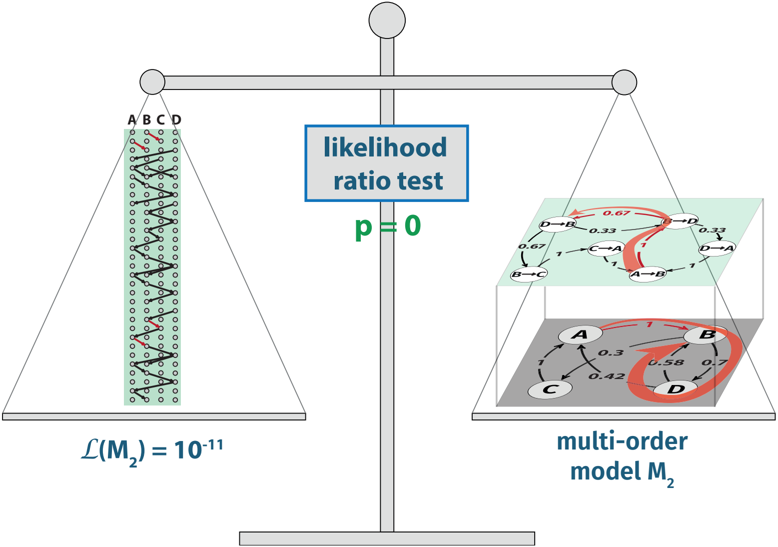
Higher-order network analysis

helps us to ...

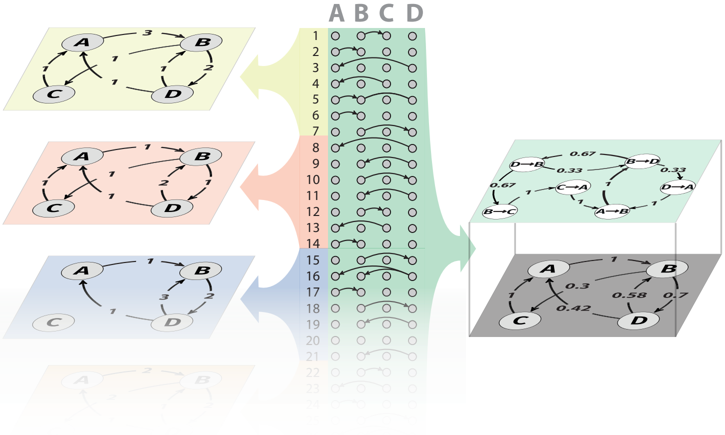
- understand **flow and spreading processes** in dynamic networks
- **detect temporal-topological clusters** in time series data
- **identify influential individuals** in time-resolved social networks
- **recognize anomalies** in temporal and sequential data

$\rightarrow$ **live coding sessions**

Learning optimal models



Temporal network analysis



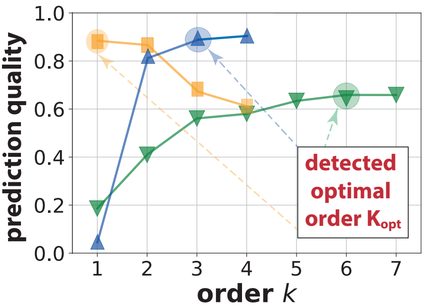
time-slice graphs

time-aggregated time slices
discard information on ordering

today's tutorial

optimal **multi-order graph**
summarisation of **causal topologies**

Optimal models!

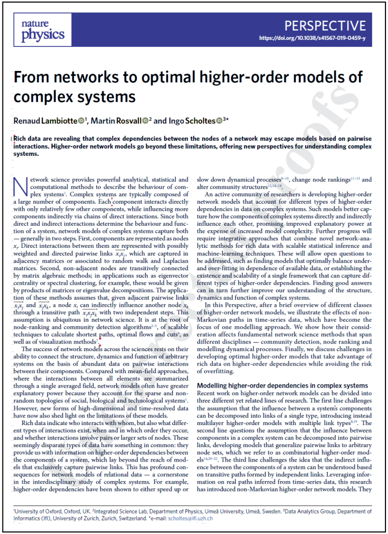


- time-stamped network data from **technical and social systems**
- how well do centralities in “learned” model predict ground truth node importance?

conclusion

models with detected **optimal order yield best prediction** → I Scholtes, SIGKDD 2017

From networks to optimal models



nature physics

PERSPECTIVE

https://doi.org/10.1038/s41567-019-0459-y

From networks to optimal higher-order models of complex systems

Renaud Lambiotte[1], Martin Rosvall[2] and Ingo Scholtes[3]*

Rich data are revealing that complex dependencies between the nodes of a network may escape models based on pairwise interactions. Higher-order network models go beyond these limitations, offering new perspectives for understanding complex systems.

Network science provides powerful analytical, statistical and computational methods to describe the behaviour of complex systems[1]. Complex systems are typically composed of a large number of components. Each component interacts directly with only relatively few other components, while influencing more components indirectly via chains of direct interactions. Since both direct and indirect interactions determine the behaviour and function of a system, network models of complex systems capture both — generally in two steps. First, components are represented as nodes x_i. Direct interactions between them are represented with possibly weighted and directed pairwise links $\overrightarrow{x_i x_j}$, which are captured in adjacency matrices or associated to random walk and Laplacian matrices. Second, non-adjacent nodes are transitively connected by matrix algebraic methods; in applications such as eigenvector centrality or spectral clustering, for example, these would be given by products of matrices or eigenvalue decompositions. The application of these methods assumes that, given adjacent pairwise links $\overrightarrow{x_i x_j}$ and $\overrightarrow{x_j x_k}$, a node x_i can indirectly influence another node x_k through a transitive path $\overrightarrow{x_i x_j x_k}$ with two independent steps. This assumption is ubiquitous in network science. It is at the root of node-ranking and community detection algorithms[2–5], of scalable techniques to calculate shortest paths, optimal flows and cuts[6], as well as of visualization methods[7].

The success of network models across the sciences rests on their ability to connect the structure, dynamics and function of arbitrary systems on the basis of abundant data on pairwise interactions between their components. Compared with mean-field approaches, where the interactions between all elements are summarized through a single averaged field, network models often have greater explanatory power because they account for the sparse and non-random topologies of social, biological and technological systems[1]. However, new forms of high-dimensional and time-resolved data have now also shed light on the limitations of these models.

Rich data indicate who interacts with whom, but also what different types of interactions exist, when and in which order they occur, and whether interactions involve pairs or larger sets of nodes. These seemingly disparate types of data have something in common: they provide us with information on higher-order dependencies between the components of a system, which lay beyond the reach of models that exclusively capture pairwise links. This has profound consequences for network models of relational data — a cornerstone in the interdisciplinary study of complex systems. For example, higher-order dependencies have been shown to either speed up or slow down dynamical processes[8–10], change node rankings[11–13] and alter community structures[12,14–18].

An active community of researchers is developing higher-order network models that account for different types of higher-order dependencies in data on complex systems. Such models better capture how the components of complex systems directly and indirectly influence each other, promising improved explanatory power at the expense of increased model complexity. Further progress will require integrative approaches that combine novel network-analytic methods for rich data with scalable statistical inference and machine-learning techniques. These will allow open questions to be addressed, such as finding models that optimally balance under- and over-fitting in dependence of available data, or establishing the existence and scalability of a single framework that can capture different types of higher-order dependencies. Finding good answers can in turn further improve our understanding of the structure, dynamics and function of complex systems.

In this Perspective, after a brief overview of different classes of higher-order network models, we illustrate the effects of non-Markovian paths in time-series data, which have become the focus of one modelling approach. We show how their consideration affects fundamental network science methods that span different disciplines — community detection, node ranking and modelling dynamical processes. Finally, we discuss challenges in developing optimal higher-order models that take advantage of rich data on higher-order dependencies while avoiding the risk of overfitting.

Modelling higher-order dependencies in complex systems

Recent work on higher-order network models can be divided into three different yet related lines of research. The first line challenges the assumption that the influence between a system's components can be decomposed into links of a single type, introducing instead multilayer higher-order models with multiple link types[8,19]. The second line questions the assumption that the influence between components in a complex system can be decomposed into pairwise links, developing models that generalize pairwise links to arbitrary node sets, which we refer to as combinatorial higher-order models[14,20–22]. The third line challenges the idea that the indirect influence between the components of a system can be understood based on transitive paths formed by independent links. Leveraging information on real paths inferred from time-series data, this research has introduced non-Markovian higher-order network models. They

[1]University of Oxford, Oxford, UK. [2]Integrated Science Lab, Department of Physics, Umeå University, Umeå, Sweden. [3]Data Analytics Group, Department of Informatics (IfI), University of Zurich, Zurich, Switzerland. *e-mail: scholtes@ifi.uzh.ch

Nature Physics, March 2019

temporal data on networks

PATHPY

http://www.pathpy.net

optimal model of causal topology