Understanding Complex Systems

From Networks to Optimal Higher-Order Models

Ingo Scholtes

Data Analytics Group
University of Zurich
Switzerland

scholtes@ifi.uzh.ch

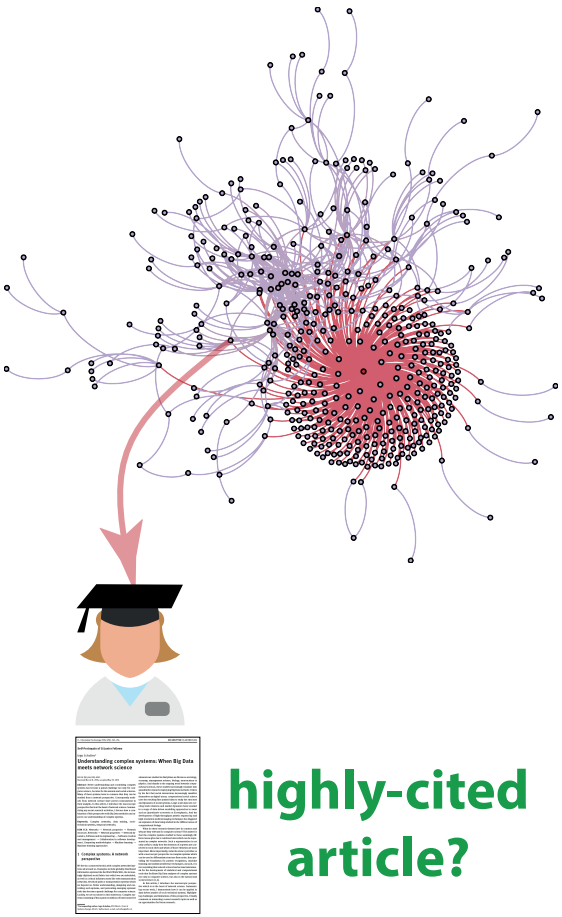
Introductory Lecture
Hands-On Tutorial
Complexity Science Hub Vienna

2018/09/05

Data analytics in complex systems

- **Modelling Social Organisations**
  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. evolution of complex software architectures
     Int. Conf. on Modularity, April 2014
  6. team structures and team performance
     Empirical Software Engineering, April 2016
- **Data Analytics for Complex Relational Data**
  7. network inference from noisy sensor data
     SocInfo, August 2017
  8. model order reduction for relational data
     under review, April 2018
  9. network analytics for time series data
     SIGKDD, August 2017

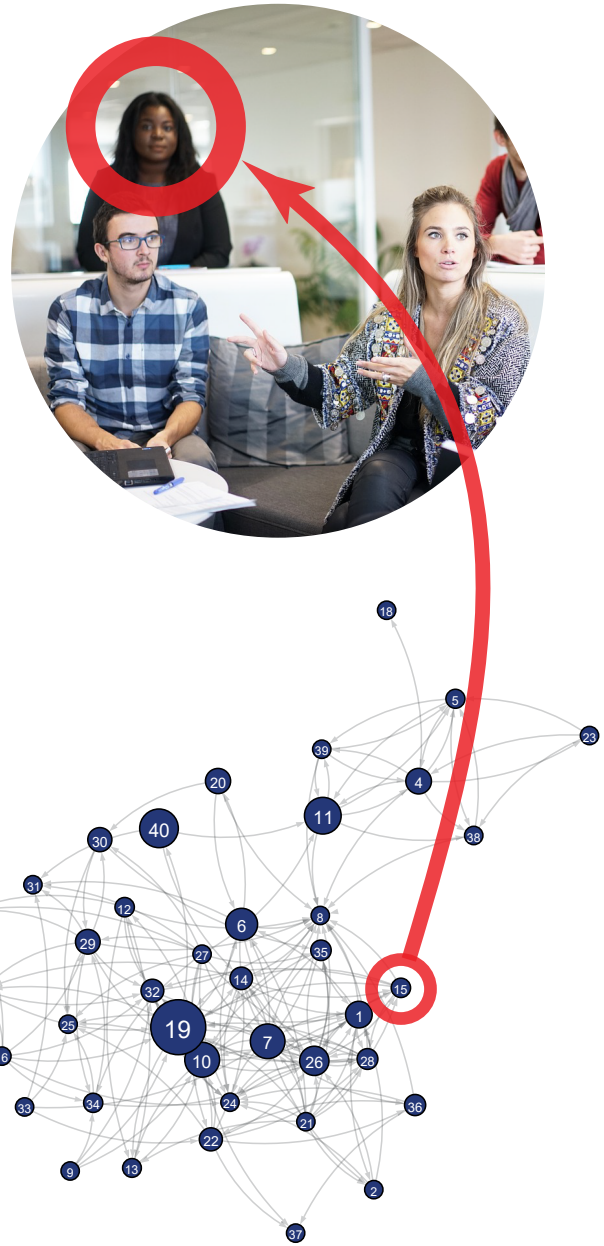
scientific collaborations



highly-cited article?

Network models?

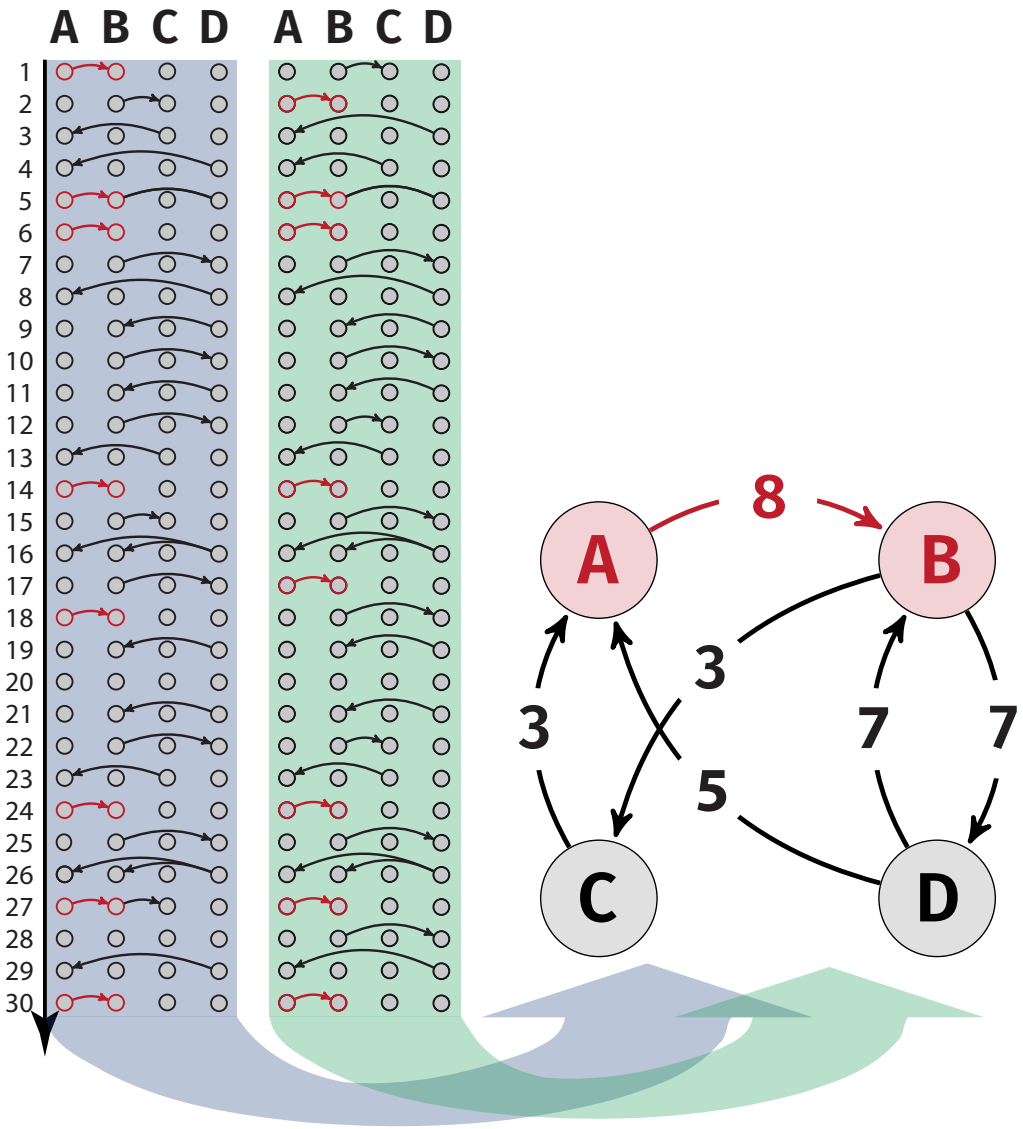
from → to	when
10 → 15	14:12:21
43 → 38	14:12:47
22 → 57	14:12:49
19 → 35	14:13:05
2 → 25	14:13:12
48 → 31	14:13:29
30 → 21	14:13:33
8 → 7	14:14:01
55 → 17	14:14:17
20 → 27	14:14:57
3 → 40	14:15:03
51 → 5	14:15:07
4 → 24	14:16:34
56 → 5	14:16:35
3 → 57	14:18:24
31 → 11	14:18:28
1 → 6	14:18:32
40 → 3	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 graph representations**
- **limits network analysis** of social, technical, and biological systems

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

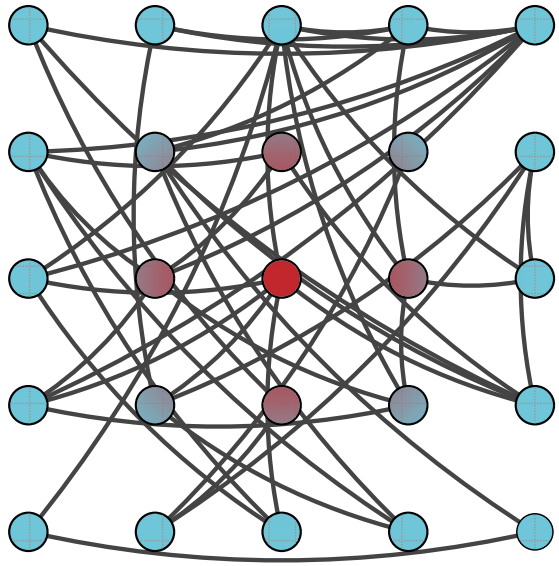
Ordering matters!

- 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

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

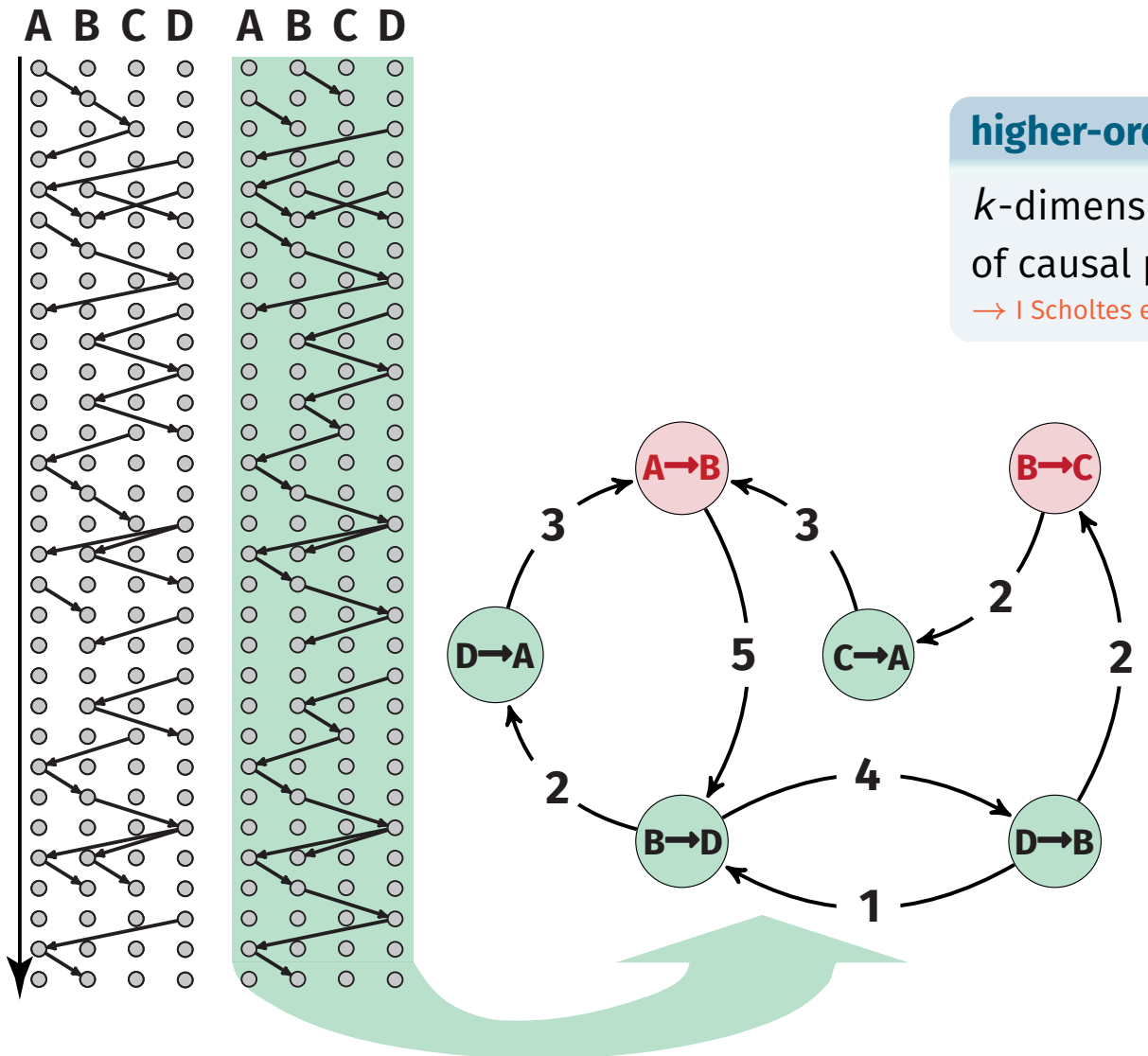


$$\frac{d\psi}{dt} = -C\mathcal{L}\psi$$

$$\mathcal{L} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ 25 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & \dots & 25 \end{array} \\ \begin{bmatrix} \mathbf{3} & 0 & 0 & 0 & \dots & 0 \\ 0 & \mathbf{2} & 0 & -1 & \dots & 0 \\ 0 & 0 & \mathbf{7} & -1 & \dots & 0 \\ 0 & -1 & -1 & \mathbf{4} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \mathbf{2} \end{bmatrix} \end{array}$$

$$\lambda_1(\mathcal{L}), \lambda_2(\mathcal{L}), \dots$$

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

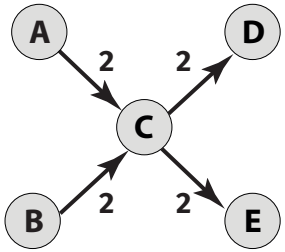
Time-aware node ranking

centrality measures

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

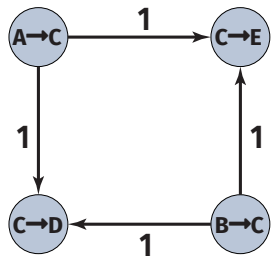
temporal-topological centrality

- 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



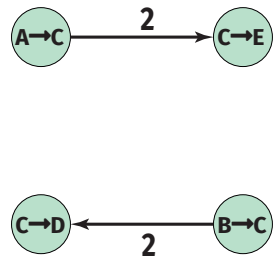
first-order node centralities

betweenness(**C**) = 4



second-order model 1

betweenness(**C**) = 4



second-order model 2

betweenness(**C**) = 2

Time-aware graph clustering

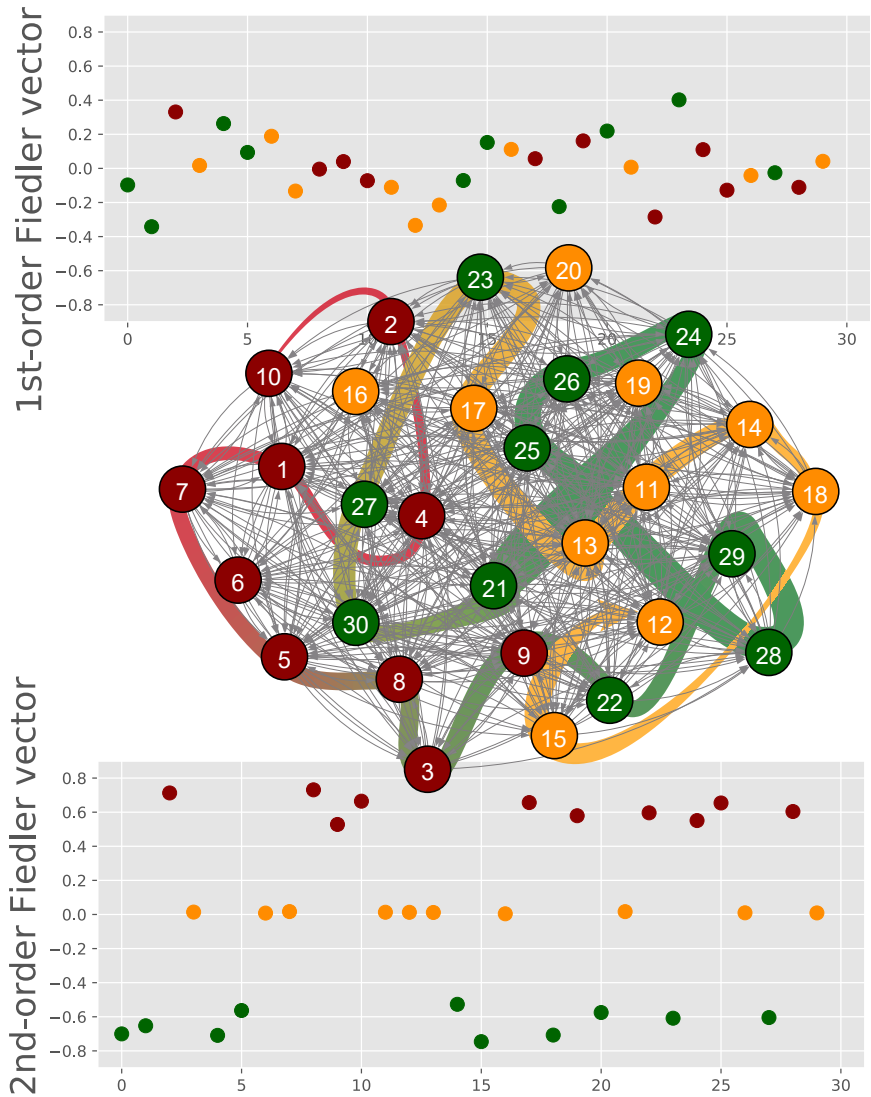
graph clustering algorithms

- fnd clusters of “well-connected” nodes
  - **modularity optimisation**
  - **clique percolation**
  - **flow compression**
  - **spectral clustering**
- but: 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



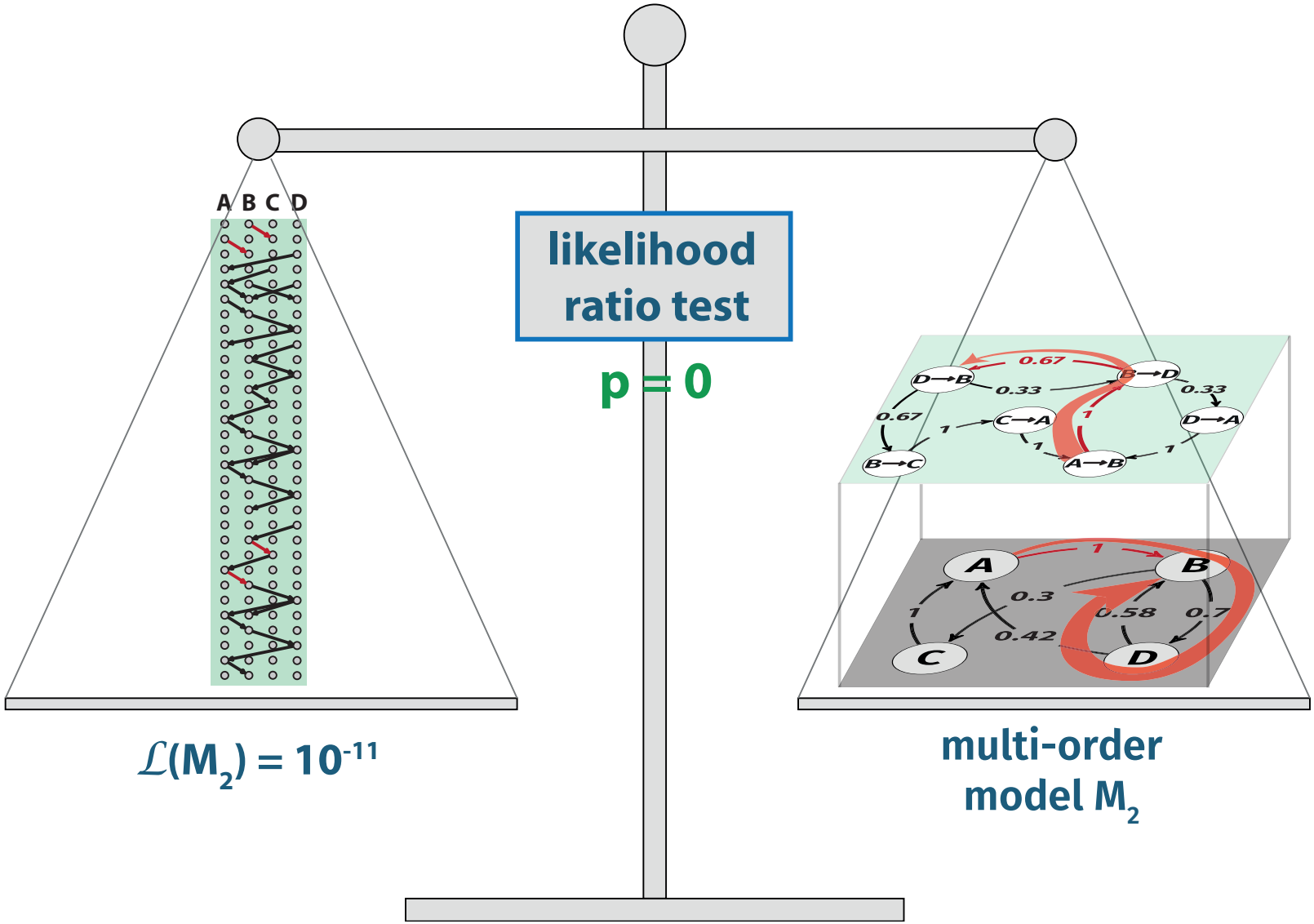
Higher-order network analytics

helps us to ...

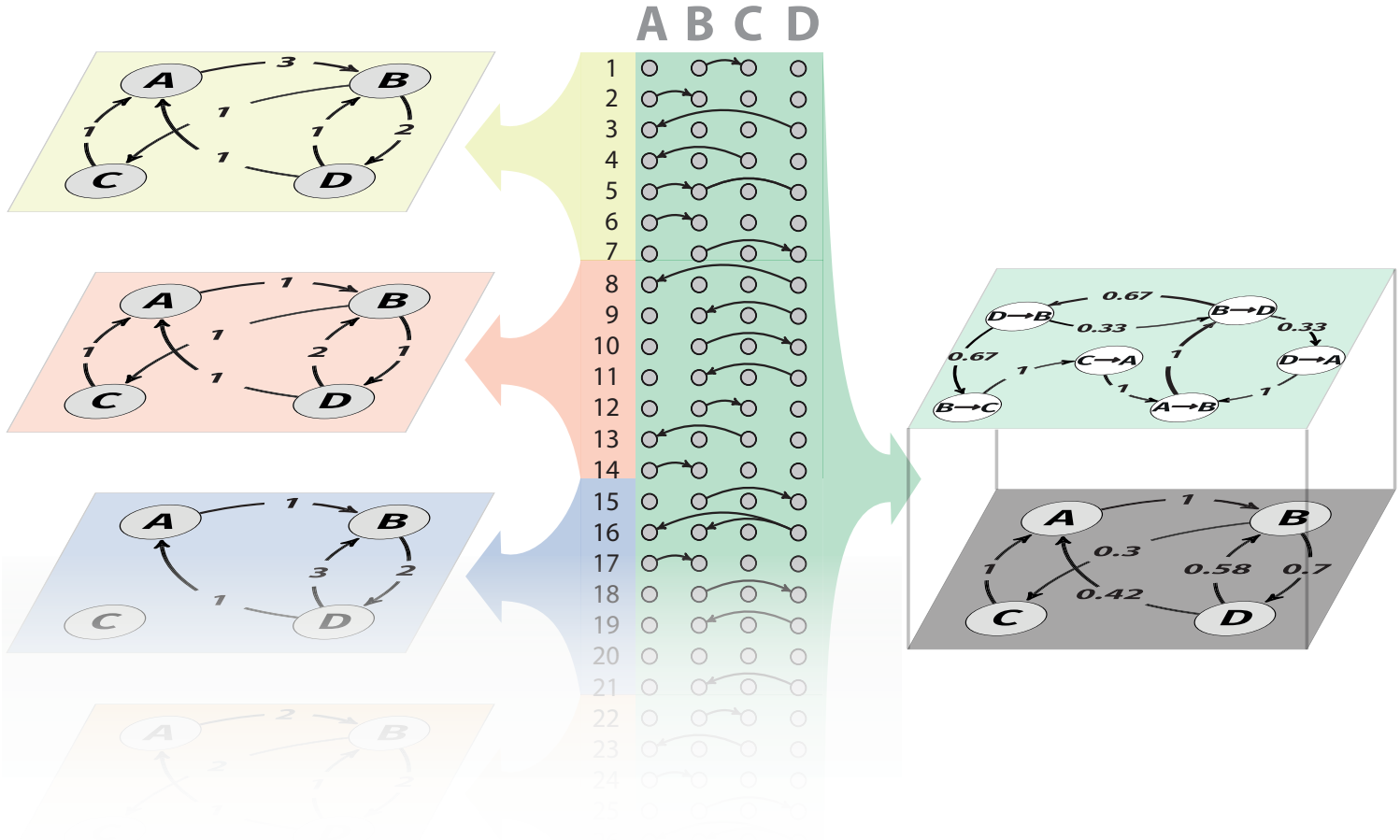
- analytically understand **flow and spreading processes** in dynamic network topologies
- **detect temporal-topological clusters** in time series data on networks
- **recognize anomalies** in temporal and sequential data
- **identify influential individuals** in time-resolved social networks

→ **live coding sessions**

Representation learning



Temporal network analysis



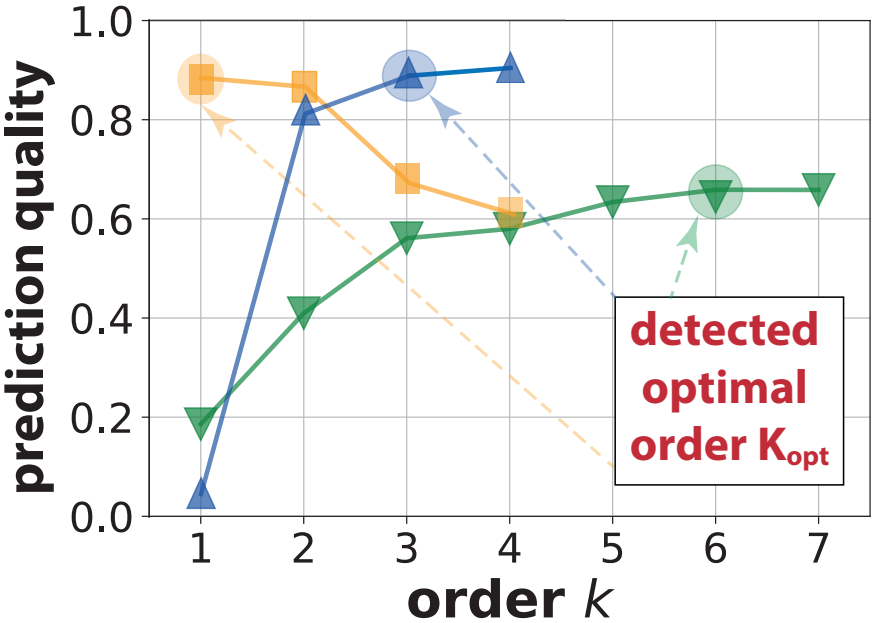
time-slice graphs

time-aggregated time slices
discard information on ordering

our perspective

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

Understanding Complex Systems:
From Networks to Optimal Higher-Order Models

Renaud Lambiotte[†], Martin Rosvall[⋆], Ingo Scholtes[‡]
[†] University of Oxford, United Kingdom
[⋆] Integrated Science Lab, Umeå University, Sweden
[‡] Chair of Systems Design, ETH Zürich, Switzerland

Abstract

To better understand the structure and function of complex systems, researchers often represent direct interactions between components in complex systems with networks, assuming that indirect influence between distant components can be modelled by paths. Such network models assume that actual paths are memoryless. That is, the way a path continues as it passes through a node does not depend on where it came from. Recent studies of data on actual paths in complex systems question this assumption and instead indicate that memory in paths does have considerable impact on central methods in network science. A growing research community working with so-called higher-order network models addresses this issue, seeking to take advantage of information that conventional network representations disregard. Here we summarise the progress in this area and outline remaining challenges calling for more research.

A long-standing goal of statistical physics is to understand emergent phenomena in complex systems that consist of a large number of interacting components. Such systems not only occur in condensed matter physics, but they are also widespread in other disciplines, and physicists have been able to contribute to a better understanding of biological, social, economic, and technological systems. A salient feature of complex systems is that all system components can influence each other, either directly or indirectly. Systems for which the topology of these interactions is unknown are often studied using mean-field approaches, which summarise interactions between all elements with a single averaged field. Over the past few decades, a surge of data has demonstrated that complex systems in the real world exhibit sparse and complex topologies in which few components directly interact with each other, while most components indirectly influence each other via sequences of multiple direct interactions [1]. Such systems can be conveniently represented as *graphs* or *networks*, where nodes x_i represent the components of a system and *links* $\overrightarrow{x_i x_j}$ capture the topology of direct pairwise interactions. The indirect influence between two components x_0 and x_n can be studied based on sequences of direct interactions $\overrightarrow{x_0 x_1}, \overrightarrow{x_1 x_2}, \ldots, \overrightarrow{x_{n-1} x_n}$ that mediate the influence between x_0 and x_n via a *path* $\overrightarrow{x_0 \ldots x_n}$.

Building on this abstraction, *network science* has developed methods that help us to better understand the structure and function of complex systems. The success and popularity of these network science methods across disciplines rest on their broad applicability to relational data that capture pairwise interactions. However, the analysis of such data based on linear transformations, algebraic methods, and Markovian models of dynamical processes [2, 3] also makes an important assumption, namely that the *paths* by which a system's components *indirectly* influence each other can be understood based on the *transitive closure* of direct, pairwise interactions. That is, most network science methods rest on the assumption that the existence of direct interactions $\overrightarrow{x_0 x_1}$ and $\overrightarrow{x_1 x_2}$ implies that x_0 can indirectly influence x_2 via a transitive path $\overrightarrow{x_0 x_1 x_2}$. Notably, methods based on the composition of linear transformations that capture network topologies implicitly introduce this fundamental hypothesis about indirect influence in a complex system. Examples include algebraic and spectral methods based on eigenvalues, products, and powers of adjacency matrices and Laplacians;

arXiv:1806.05977

temporal data on networks

PATHPY

`http://www.pathpy.net`

optimal model of causal topology