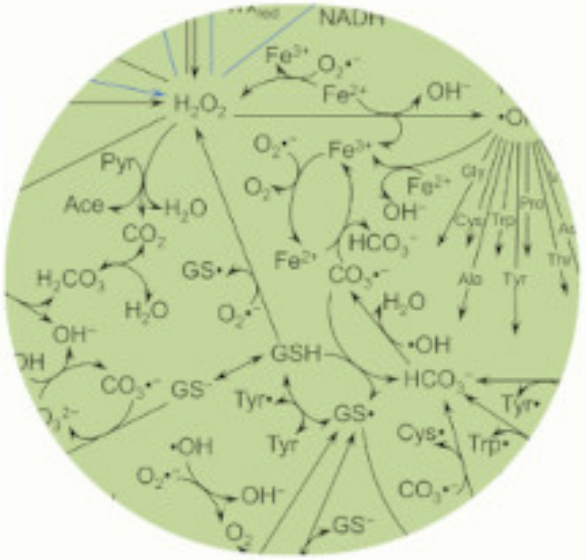
Discovering Causality in Event Time-Series

bionetgen/
bionetgen

msneddon/**nfsim**

A general-purpose, stochastic, biochemical reaction simulator for large reaction networks



Python

BASH

Pavel Loskot
pavelloskot@intl.zju.edu.cn

ZJU-UIUC INSTITUTE
Zhejiang University-University of Illinois at Urbana-Champaign Institute
浙江大学伊利诺伊大学厄巴纳香槟校区联合学院

The Seventh International Conference on Advances in Signal, Image and Video Processing
SIGNAL 2022
May 22, 2022 to May 26, 2022 - Venice, Italy

IARIA

About Me

Pavel Loskot joined the ZJU-UIUC Institute as Associate Professor in January 2021. He received PhD in Wireless Communications from University of Alberta, Canada, and MSc and BSc in Radioelectronics and Biomedical Electronics, respectively, from the Czech Technical University of Prague. He is Senior Member of the IEEE, Fellow of the HEA, and the Recognized Research Supervisor of the UKCGE in the UK.

In the past 25 years, he was involved in numerous industrial and academic collaborative projects in the Czech Republic, Finland, Canada, UK, Turkey, and China. These projects concerned mainly wireless and optical telecommunication networks, but also genetic circuits, air transport services, and renewable energy systems. This experience allowed him to truly understand the interdisciplinary workings, and crossing the disciplines boundaries.

His current research focuses on statistical signal processing, classical machine learning, and importing methods from Telecommunication Engineering and Computer Science to model and analyze systems more efficiently and with greater information power.

OBJECTIVES

- describe dynamic systems by their event history
- discover causality among events being categorical random variables
- implement automated causal event discovery for interpretable stochastic simulations of biochemical reaction networks

OUTLINE

- events in state-space model of dynamic systems
- causality as conditional probabilities
- modified matrix profile analysis
- simulation software and numerical examples

Introduction

Causality

- key in scientific reasoning: hypothesis discovery and testing
- key in engineering reasoning: prescriptive statistics
- statistical associations are neither necessary nor sufficient to infer causality

Literature

- discovering causality in time-series data
  → structural causal models (SCMs)
  → Granger and intervention causality, average causal effect (ACE)
- matrix profile analysis of time-series data
  → discovering motifs and discords, identifying changes in statistics

New problem

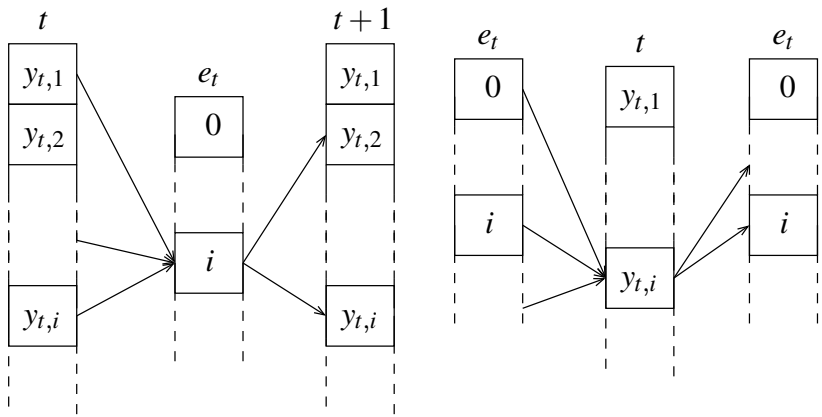
- causality in categorical time-series
  → event history of a dynamic system
- causality between events newly defined here as
  → nearly certain and uncertain conditional event sequences
- a versatile causal analysis framework, which can be fully automated
  → event sets and multi-sets and their distances and equivalences
  → modified matrix profile

System Model

Dynamic system

- state-space representation

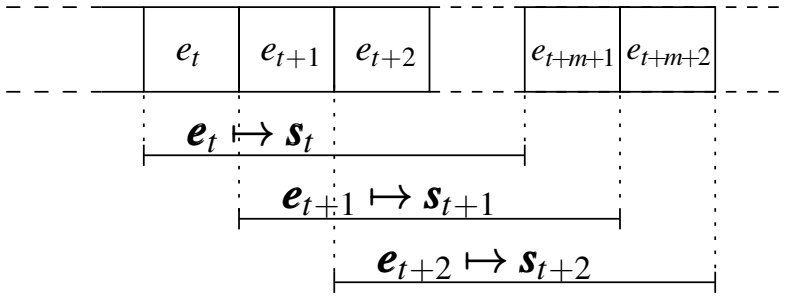
$$\begin{aligned} \boldsymbol{z}_t &= e_t(\boldsymbol{z}_{t-1}, \boldsymbol{z}_{t-2}, \ldots) \\ \boldsymbol{y}_t &= O(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \ldots) \end{aligned}$$



- assumptions

(1) system is memoryless:
$\boldsymbol{z}_{t+1} = e_t(\boldsymbol{z}_t)$ and $\boldsymbol{y}_t = O(\boldsymbol{z}_t)$

(2) observations are perfect: $\boldsymbol{y}_t = \boldsymbol{z}_t$



Theorem Given observations $\boldsymbol{y}_t$ at time t, the ordering of $(m+1) > 0$ events in sequence, $(e_t, e_{t+1}, \ldots, e_{t+m})$, does not affect the observation $\boldsymbol{y}_{t+m}$ at time $(t+m)$.

Consequently

- event ordering in shorter sequences is irrelevant
- event sequences can be converted to sets or multi-sets

Analysis of Event Time-Series

Process

- sliding window partitioning to create sequences $\boldsymbol{e}_t$ of N events each
- convert sequences $\boldsymbol{e}_t$ to (multi-) sets $\boldsymbol{s}_t$
- compute empirical conditional probabilities $\Pr\left(\boldsymbol{e}_j|\boldsymbol{e}_i\right)$ or $\Pr\left(\boldsymbol{s}_j|\boldsymbol{s}_i\right)$
  $\rightarrow$ in general, $\Pr\left(\boldsymbol{e}_j|\boldsymbol{e}_i\right) \neq \Pr\left(\boldsymbol{e}_i|\boldsymbol{e}_j\right)$

Definition 1 The event sequences $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, $j > i$, have a cause-effect relationship, provided that their conditional probability, $\Pr\left(\boldsymbol{e}_j|\boldsymbol{e}_i\right) \rightarrow 1$.

Definition 2 The event sequences $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, $j > i$, have no cause-effect relationship, provided that $\Pr\left(\boldsymbol{e}_j|\boldsymbol{e}_i\right) \rightarrow 0$ and $\Pr\left(\boldsymbol{e}_i|\boldsymbol{e}_j\right) \rightarrow 0$.

Practical implementation

- split every $\boldsymbol{e}_t = \boldsymbol{e}_i \cup \boldsymbol{e}_j$, so that $\boldsymbol{e}_i \cap \boldsymbol{e}_j = \emptyset$, $|\boldsymbol{e}_i| = N_1$, $|\boldsymbol{e}_j| = N_2$, and $N_1 + N_2 = N$
  $\rightarrow$ referred to as left and right sub-sequences

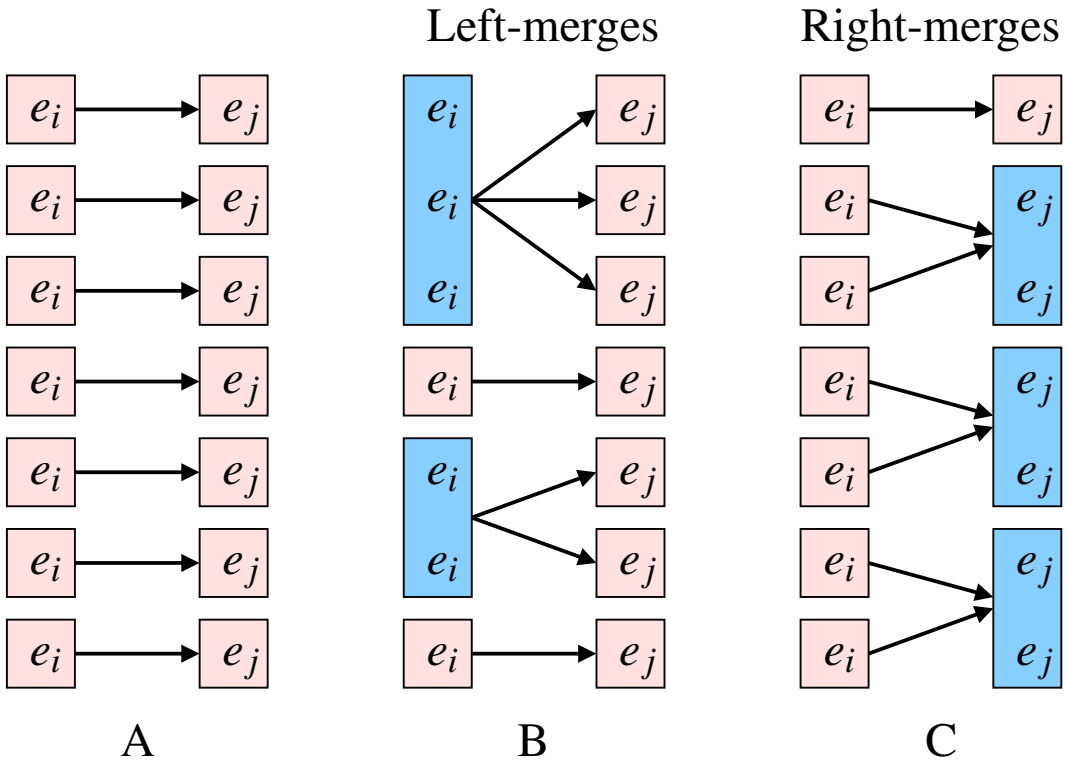
Event equivalence The event (sub-) sequences $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ are said to be equivalent, provided that their distance, $d(\boldsymbol{e}_i, \boldsymbol{e}_j) = 0$.
$\rightarrow$ enables to make $\Pr\left(\boldsymbol{e}_j|\boldsymbol{e}_i\right)$ closer to 1

Analysis of Event Time-Series (2)

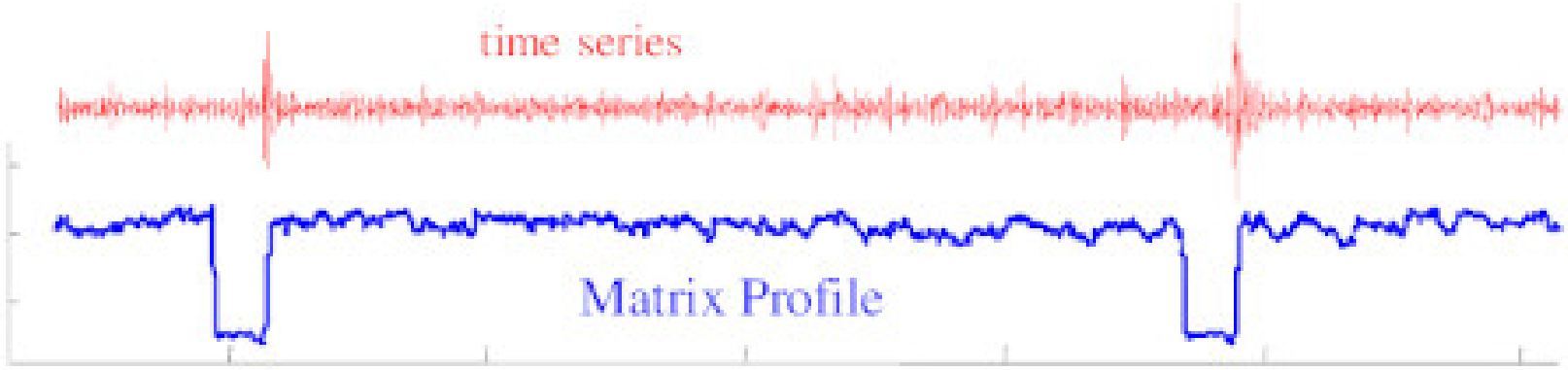
Distance between two event sequences

$$
\begin{aligned}
d(\boldsymbol{e}_i, \boldsymbol{e}_j) &= d_0 - |\boldsymbol{s}_i \cup \boldsymbol{s}_j| \\
d(\boldsymbol{e}_i, \boldsymbol{e}_j) &= d_0 - |\boldsymbol{s}_i \cap \boldsymbol{s}_j| \\
d(\boldsymbol{e}_i, \boldsymbol{e}_j) &= d_0 - (|\boldsymbol{s}_i| + |\boldsymbol{s}_j|) \\
d(\boldsymbol{e}_i, \boldsymbol{e}_j) &= d_0 - \max(|\boldsymbol{s}_i|, |\boldsymbol{s}_j|) \\
d(\boldsymbol{e}_i, \boldsymbol{e}_j) &= \max(|\boldsymbol{s}_i|, |\boldsymbol{s}_j|) - \min(|\boldsymbol{s}_i|, |\boldsymbol{s}_j|) \\
d(\boldsymbol{e}_i, \boldsymbol{e}_j) &= \min(|\boldsymbol{s}_i \setminus \boldsymbol{s}_j|, |\boldsymbol{s}_j \setminus \boldsymbol{s}_i|)
\end{aligned}
$$

Merging left/right event sub-sequences



Matrix Profile Analysis



Canonical matrix profile

- given sliding window sub-sequences of length m, shows the minimum distance from any other distinct sub-sequence
  $\rightarrow$ the profile strongly affected by actual choice of m
- efficient implementations exist in different languages
  $\rightarrow$ e.g. stumpy in Python

Modified matrix profile

- for categorical event sequences, already defined various distance metrics
  $\rightarrow$ can be generalized to any other pairwise metrics
- can consider both the minimum/maximum distance as well as its multiplicity

Numerical Examples

Biochemical reaction network

- a dynamic system described by Chemical Master Equation (CME)
- stochastic simulations are often used to solve CME
  → time evolution of chemical species counts
- chemical reactions (i.e., events) are occurring sequentially at random
- the goal is to analyze the event history (time-series)
  → it is readily available in simulations by recording all reaction events
  → this is an alternative approach to analyzing chemical species counts

Procedure

- select a specific biochemical system to simulate
- obtain the SBML model of this system and simulate it
  → BioNetGen and NFsim open source software
- record and process the reaction events
  → Python and Bash scripts
- the whole procedure can be largely automated
  → generated 160 diverse plots for 12 experiments across 9 biochemical models in less than 1 hour with minimum manual intervention

Numerical Examples (2)

Simulated biochemical model

- antigen receptor signaling regulating the activity and fate of the B-cells
- 32 molecule types, 158 reaction rules, 129 model parameters
- full extracted model: 1,124 chemical species and 24,390 chemical reactions
- 100 s of simulation time resulted in 3,634,390 reaction events
  → can be divided in blocks, e.g. once per 1 s

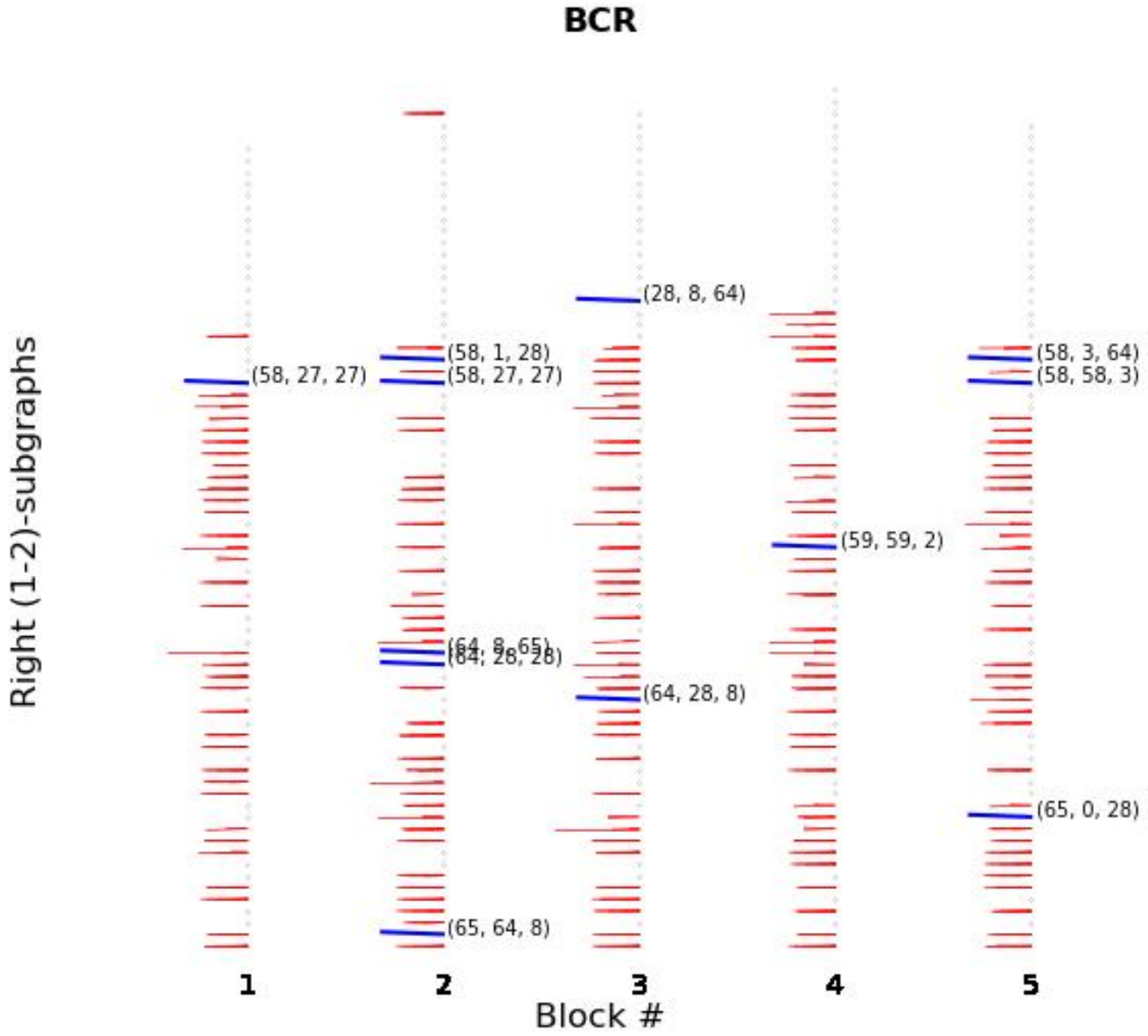
NFsim modifications

- upgraded source code last updated in 2011 to more recent C++ compiler
  → resolved compiling errors, removed most warnings
- added event recording functionality
  → written in text or binary file

BioNetGen

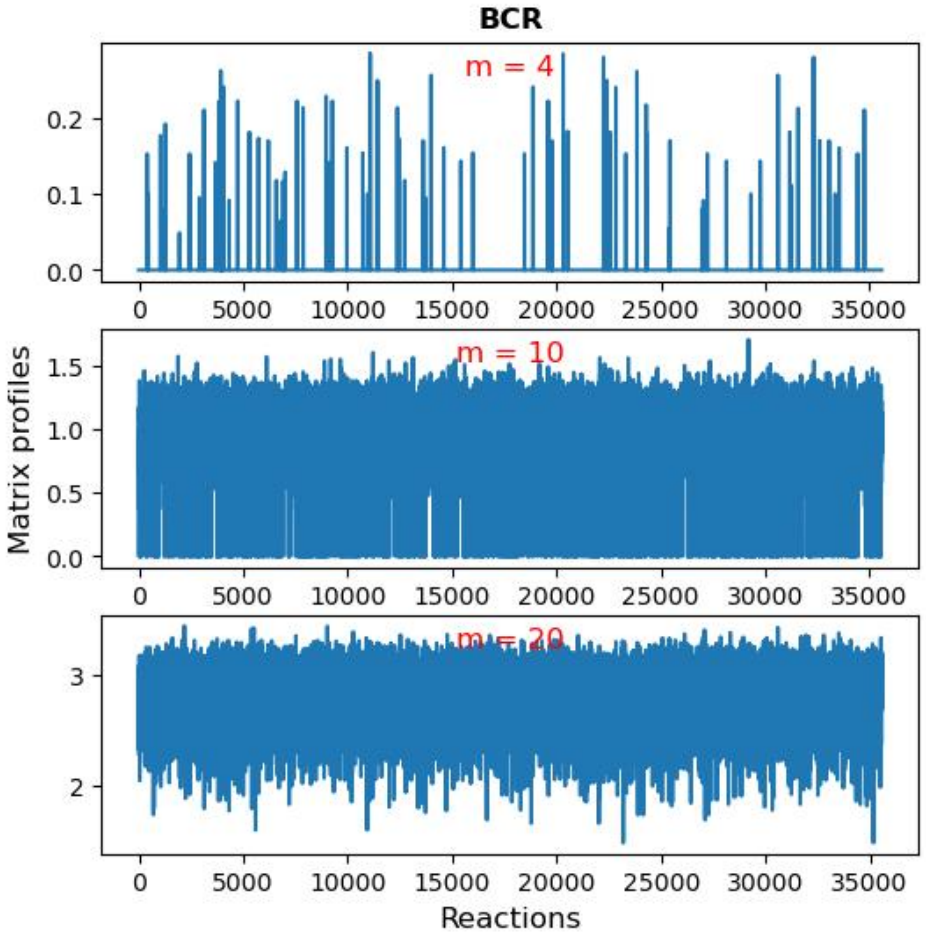
- biochemical model described in human-friendly bngl text file
- a Perl script in BioNetGen processes bngl file into SBML file
- SBML, the model description in XML, is simulated in NFsim
  → BioNetGen acts as a user-friendly interface for NFsim

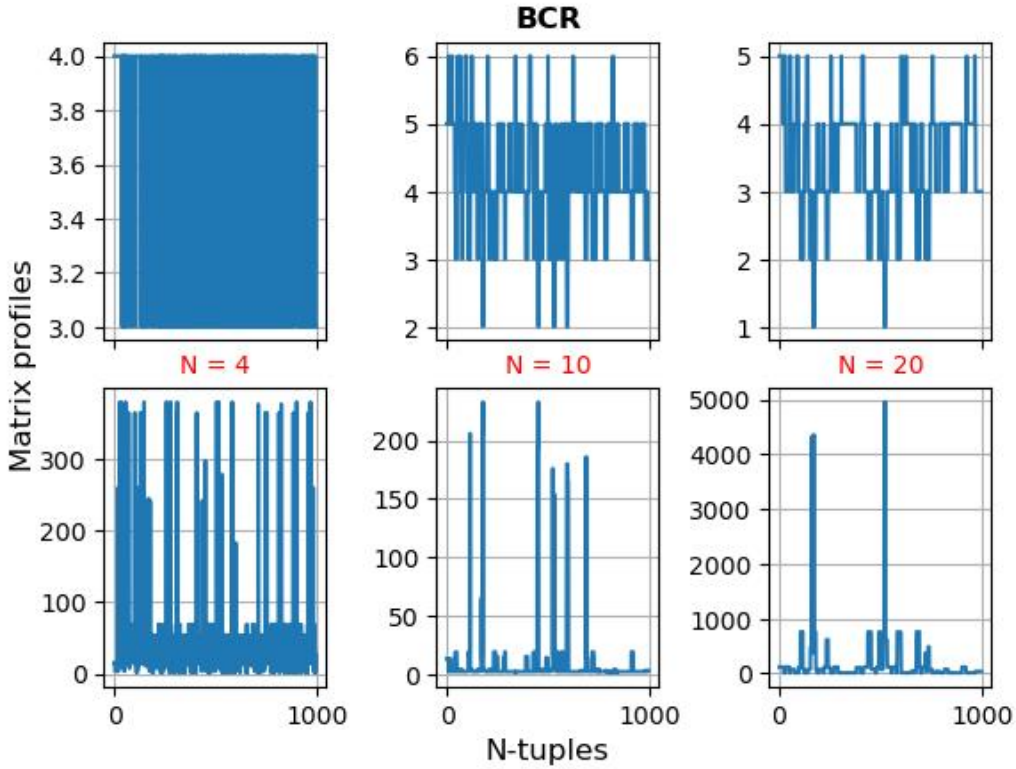
Numerical Examples (3)



- $N = 3$ with $|\boldsymbol{e}_j| = 2$ and $|\boldsymbol{e}_i| = 1$, and assuming $\Pr\left(\boldsymbol{e}_i|\boldsymbol{e}_j\right) \equiv P_{j,i}$
- Black: $P_{j,i} < 0.1$, Red: $P_{j,i} < 0.9$, Blue: $P_{j,i} \geq 0.9$

Numerical Examples (4)





- Left: canonical matrix profile with minimum distances $d(\boldsymbol{e}_i, \boldsymbol{e}_j) = d_0 - |\boldsymbol{s}_i \cup \boldsymbol{s}_j|$
- Right: modified matrix profile with maximum distances $d(\boldsymbol{e}_i, \boldsymbol{e}_j) = \min(|\boldsymbol{s}_i \setminus \boldsymbol{s}_j|, |\boldsymbol{s}_j \setminus \boldsymbol{s}_i|)$

Conclusion

Key points

- behavior of dynamic systems accurately described by the events history
- ordering of events locally unimportant, can assume (multi-) sets instead
- causality defined here as conditional nearly certain or uncertain events
- distance metrics between event sequences used for matrix profile analysis

Future work

- causality of sub-sampled event sequences
- causality of longer event sequences
- automated interpretation of discovered causally related events

Thank you!

pavelloskot@intl.zju.edu.cn