ANOMALY DETECTION IN TEMPORAL GRAPH DATA: AN ITERATIVE TENSOR DECOMPOSITION AND MASKING APPROACH

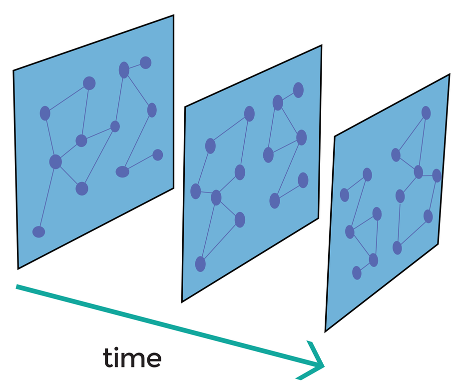
Anna Sapienza[1,2], André Panisson[2], Joseph Wu[3], Laetitia Gauvin[2], Ciro Cattuto[2]

[1]Polytechnic University of Turin, Turin, Italy
[2]Data Science Laboratory, ISI Foundation, Turin, Italy
[3]School of Public Health, University of Hong Kong, Hong Kong

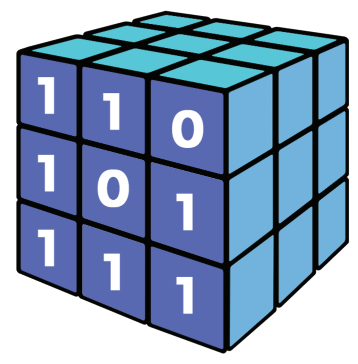
Introduction

Emerging applications and data sources in the Internet-of-things domain pose new challenges in data mining. Data generated by these new sources are often multi-relational and can be naturally represented as graphs. However, multi-dimensional graph data can suffer from anomalies that entangle temporal and topological aspects and render the detection task non-trivial. Here, a principled anomaly detection method, consisting of an iterative tensor decomposition, is developed.

Time-varying Graph



Tensor

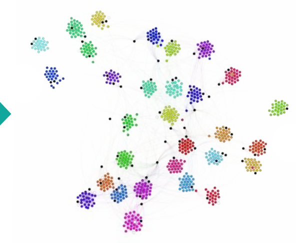


$\mathcal{T}$

Case Study

Analysis of an high-resolution social network, measured with wearable proximity sensors [1-2], describing the interactions of people in a primary school in Hong Kong

709 students
65 teachers
30 classes



N=774 nodes

10 days of observation
5 minutes of temporal resolution

S =2680 snapshots

PROBLEM

- According to the sensors wearing protocol, at the end of each day sensors were left in class in random arrangements
- The resulting spurious proximity patterns were considered as anomalies

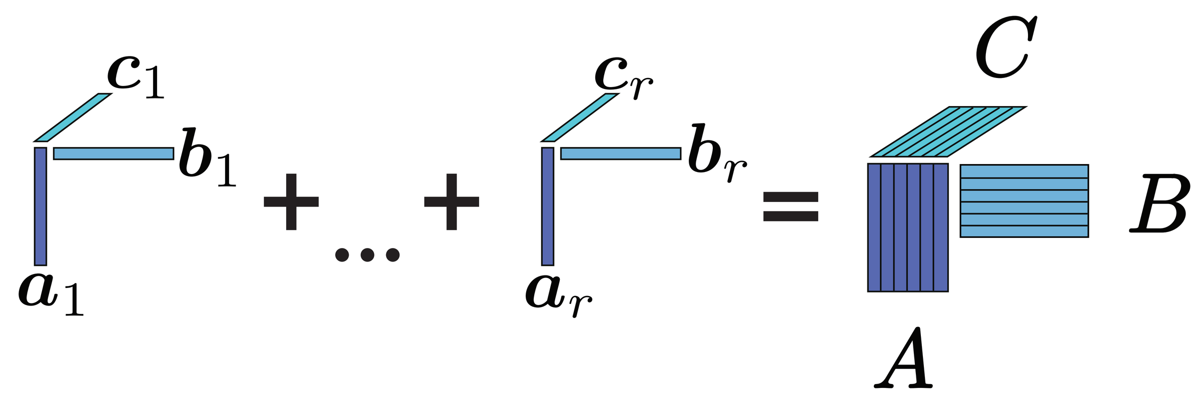
CHALLENGE

- Detect the anomalous interactions by mining the topological and temporal properties of the time-varying graph

www.sociopatterns.org

STEP 1 - NON-NEGATIVE TENSOR FACTORIZATION

1



Non-negative tensor factorization [3] approximates the tensor $\mathcal{T} \in \mathbb{R}^{N \times N \times S}$ as a sum of R rank-one tensors, called components:

$$\mathcal{T} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

- $\mathbf{a}_r, \mathbf{b}_r$: memberships of nodes to the component r
- $\mathbf{a}_r \otimes \mathbf{b}_r$: memberships of links to the component r
- $\mathbf{c}_r$: temporal activity patterns of the component r

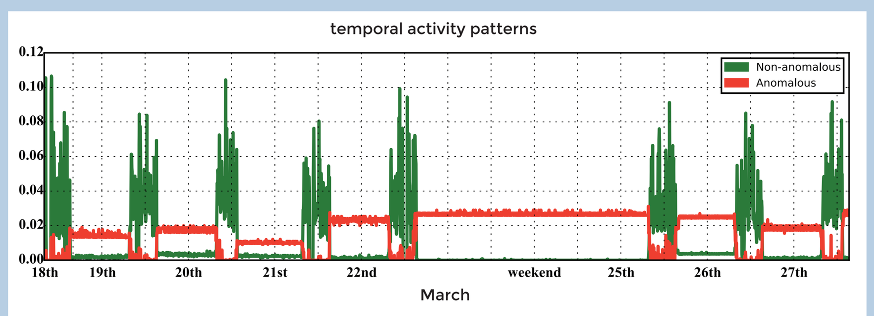
The selection of a suitable R at each iteration is guided by the Core Consistency Diagnostic [4].

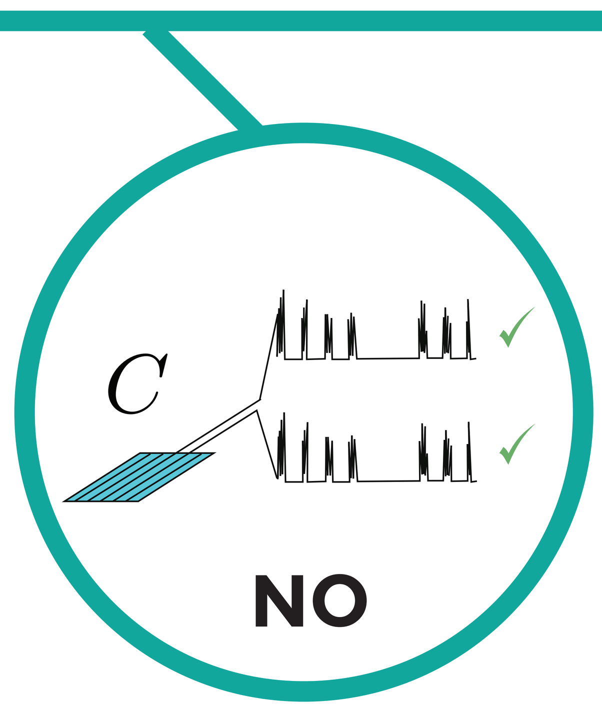
STEP 2 - CLASSIFIER

2

ARE THERE ANOMALIES ?

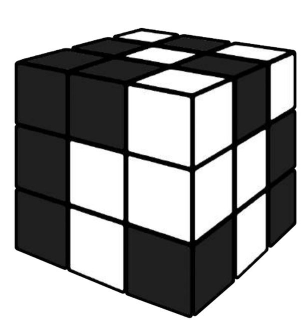
The extracted components are classified into anomalous and non-anomalous by a classifier operating on the temporal activity patterns of each component.





STEP 3 - MASK COMPUTATION

3

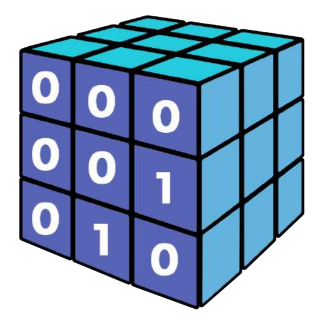


$\mathcal{M}$

The structural and temporal properties of spurious interactions are used to compute a mask $\mathcal{M}$:

- the links involved in each interaction are determined by the level of membership given by $\mathbf{a}_r \cdot \mathbf{b}_r^T$
- the occurrence times of the anomalous interactions are given by the temporal activity pattern $\mathbf{c}_r$ of the anomalous components

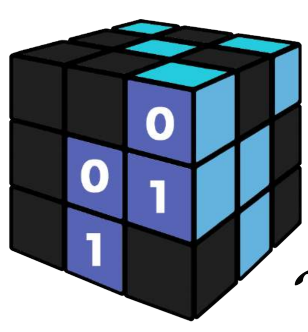
CLEANED TENSOR



$\mathcal{T}'$

STEP 4 - CLEANING PROCEDURE

4



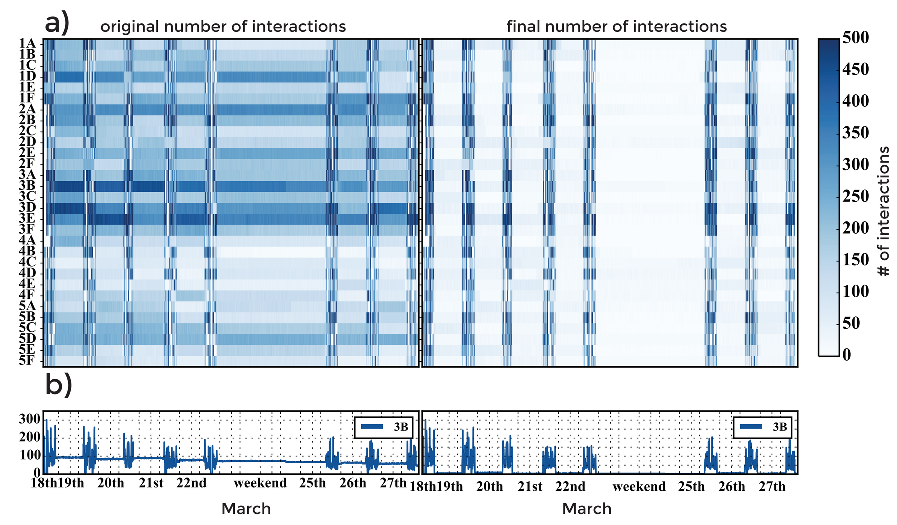
$\mathcal{T} \odot \mathcal{M}$

The mask is applied on the tensor $\mathcal{T}$ to zero out those interactions associated with anomalies and the masked tensor

$$\mathcal{T}' = \mathcal{T} \odot \mathcal{M}$$

is used as an input of the successive iteration.

RESULTS



Evolution of the number of interactions with time measured on the original and cleaned tensor for each school class.

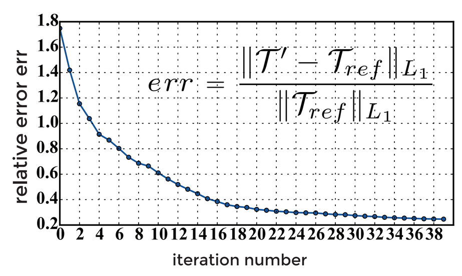
a) The class labels were used to group the interaction of nodes belonging to the same class. While interactions in the original state are distributed along the entire timeline, the cleaning procedure managed to identify and remove most of the anomalies.

b) Time series representing the evolution of the number of interactions among people belonging to one selected class, measured both on the original and cleaned tensor. On the left, it is possible to observe the great amount of interaction events recorded by sensors during the entire timeline. On the right, the corresponding time series after the application of the iterative method is shown.

VALIDATION

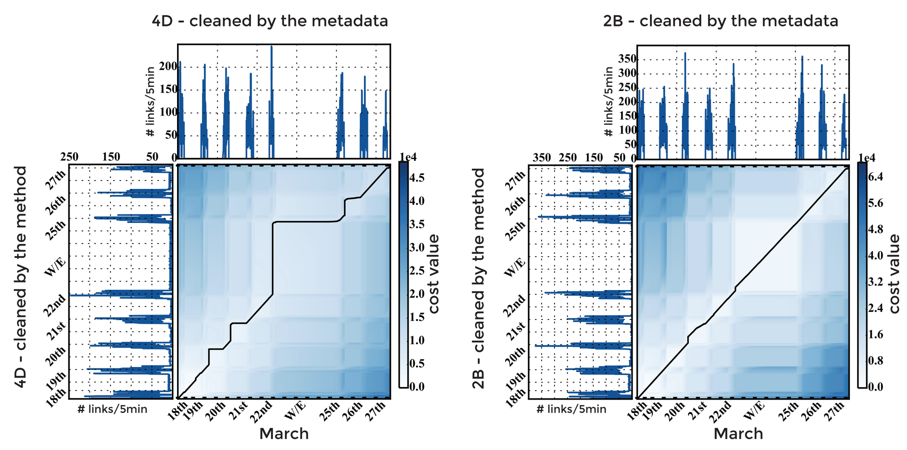
To validate the method, a reference tensor $\mathcal{T}_{ref}$ was created by using the metadata of the class membership and the school schedule.

GLOBAL LEVEL VALIDATION



	Precision	Recall	F1-score
Anomalous	0.99	0.87	0.92
Non-Anomalous	0.81	0.99	0.89
Accuracy	0.91		

CLASS LEVEL VALIDATION



DYNAMIC TIME WARPING

Example of two dynamic time warping cost matrices, in which time series measured on the cleaned tensor and the reference tensor are compared. The series are computed at the level of the classes and correspond to the evolution of the number of interactions in time. Here, classes 4D (left) and 2B (right) are shown and their Pearson coefficients are respectively 0.89 and 0.97. The lines shown in the matrices represent the optimal warping path, which is near the matrix diagonal.

Pearson coefficient	[0.89, 0.99]
p-value	$<10^{-3}$

CONCLUSION

ADVANTAGES	LIMITATIONS	FUTURE WORK
Only relies on the topological and temporal properties of the time-varying graph	Possible high computational cost	Address the limitations
Relies on standard techniques for low-rank approximations	Requires a classifier for anomalous components	Use of a higher-order tensor
Takes advantage of efficient open source library for tensor decomposition	Depends on tensor masking to remove anomalies	Use of different decomposition techniques
Succesfully tested on real-world application		Additional experiments on real-world data
Operates at the mesoscale level		

REFERENCES

[1] Cattuto, Ciro, et al. "Dynamics of person-to-person interactions from distributed RFID sensor networks." PLoS ONE 5.7 (2010): e11596.
[2] L. Gauvin, A. Panisson, and C. Cattuto, "Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach," PLoS ONE, vol. 9, no. 1, 2014.
[3] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," SIAM review, vol. 51, no. 3, pp. 455–500, 2009.
[4] R. Bro and H. A. Kiers, "A new efficient method for determining the number of components in parafac models," Journal of chemometrics, vol. 17, no. 5, pp. 274–286, 2003.

ACKNOWLEDGMENTS

This research was supported by the Lagrange Project of the ISI Foundation funded by the CRT Foundation, by the Q-ARACNE project funded by the Fondazione Compagnia di San Paolo, by the FET Multiplex Project (EU-FET-317532) funded by the European Commission, and by the Harvard Center for Communicable Disease Dynamics from the National Institute of General Medical Sciences (grant no. U54 GM088558). The funding bodies had no role in study design, data collection and analysis, preparation of the manuscript, or the decision to publish.