Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data
AALTD 2015

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

Abstract. Sensors and Internet-of-Things scenarios promise a wealth of interaction data that can be naturally represented by means of time-varying graphs. This brings forth new challenges for the identification and removal of temporal graph anomalies that entail complex correlations of topological features and activity patterns. Here we present an anomaly detection approach for temporal graph data based on an iterative tensor decomposition and masking procedure. We test this approach using high-resolution social network data from wearable sensors and show that it successfully detects anomalies due to sensor wearing time protocols.

Keywords: Data cleaning, anomaly detection, non-negative tensor factorization, high-resolution social networks, sensors, temporal networks.

1 Introduction

Emerging applications in the big data and Internet-of-Things domains pose new problems for data cleaning. Time-resolved interaction data, in particular, are especially challenging because the relational nature of the data yields anomalies that entangle temporal and topological aspects. Several studies have focused on identifying anomalous behaviors in graph-based datasets [1] and time-varying networks [2]. However, mesoscale anomalies that mimic normal behaviors are observed in empirical data and call for further research.

Here we focus on time-varying graphs [3] represented as three-mode tensors and we present an semi-supervised anomaly detection method based iterative tensor decomposition and masking. We report on the performance of this method in detecting and removing anomalies in an empirical social network dataset gathered by using wearable proximity sensors in a school.

2 Methodology

A static graph can be represented by an adjacency matrix $M \in \mathbb{R}^{N \times N}$, where $M_{ij} = 1$ if a contact between i and j occurred and $M_{ij} = 0$ otherwise. This description can be generalized to the case of a time-varying graph, by using a sequence of S consecutive adjacency matrices, that can be easily arranged as a tensor $\mathcal{T} \in \mathbb{R}^{N \times N \times S}$.

The extraction of latent structures can then be performed by following the iterative approach described below. This framework allows to carry out the data cleaning by unearthing at each iteration group behaviours of nodes having correlated activities and classifying these patterns of activities as meaningul or anomalous.

Step 1. The Non-negative Tensor Factorization [6] is used as a powerful tool to approximate the tensor $\mathcal{T}$ as a sum of R rank-one tensors $\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$, called components. In the specific case of temporal networks, $\mathbf{a}_r$ and $\mathbf{b}_r$ provide the membership of nodes to the component r, whereas $\mathbf{c}_r$ is the temporal activity pattern of the component. Moreover, it is possible to consider $\mathbf{a}_r \approx \mathbf{b}_r$ if the graph is undirected. The components can be recovered by solving an optimization problem with non-negative constraints. The minimization problem

$$\min \left\| t_{ijk} - \sum_{m=1}^{R} \sum_{n=1}^{R} \sum_{l=1}^{R} a_{im} a_{jn} c_{kl} \right\|_F^2 \quad \text{s.t. } a_{im}, a_{jn}, c_{kl} \geq 0 \tag{1}$$

is computed by the alternating non-negative least squares method [7], solved by using the block principal pivoting algorithm [8]. The selection of a suitable number R of components is guided at each iteration by the Core Consistency Diagnostic [9, 10], and performed in order to prevent overfitting.

Step 2. The extracted components are analysed in order to discriminate between those dominated by anomalous activities or meaningful behaviours. To this end, a classifier working on the temporal activity patterns of each component $\mathbf{c}_r$ was developed.

Step 3. Spurious contact patterns highlighted by the anomalous components are combined into a mask, used to clean the original tensor. The nodes involved in each of these contacts are detected by analysing the level of membership given by $\mathbf{a}_r$. The occurrence times of these contacts are given by the anomalous windows found in the temporal patterns $\mathbf{c}_r$. These windows are recovered by using a step detection algorithm based on the Otsu threshold [11].

Step 4. The mask is applied to the tensor $\mathcal{T}$ in order to erase the invalid entries. The cleaned tensor $\mathcal{T}'$ becomes then the input of the consecutive iteration in the iterative framework.

Step 5. The procedure is repeated until no component is classified as anomalous in step 2.

3 Results and Validation

The current investigation involves the analysis of a high-resolution dataset which describes the interactions of people in a primary school in Hong Kong. The school population consisted in 709 children and 65 teachers divided into 30 classes. Data were collected by using wearable proximity sensors [4, 5] over 10 consecutive days in March 2013, from Monday 18th to Thursday 27th. These sensors record spatial proximity with a resolution of 20s. As a result, a time-varying network with $N = 774$ nodes was created. The data were then aggregated over a time-window of 5min, leading to a division of the overall network in $S = 2680$ snapshots.

The protocol was as follows: the proximity of the sensors was recorded during the whole experiment duration, and the sensors were grouped in each class at the end of the school day. Hence, activity patterns composed by strong steady contacts withinh each class were observed during the school closing time. In order to clean the data, these anomalous patterns must be retrieved. A general methodology is thus developed here to deal with the anomaly detection of temporal graph-based data, and then used to perform the data cleaning of the present problem.

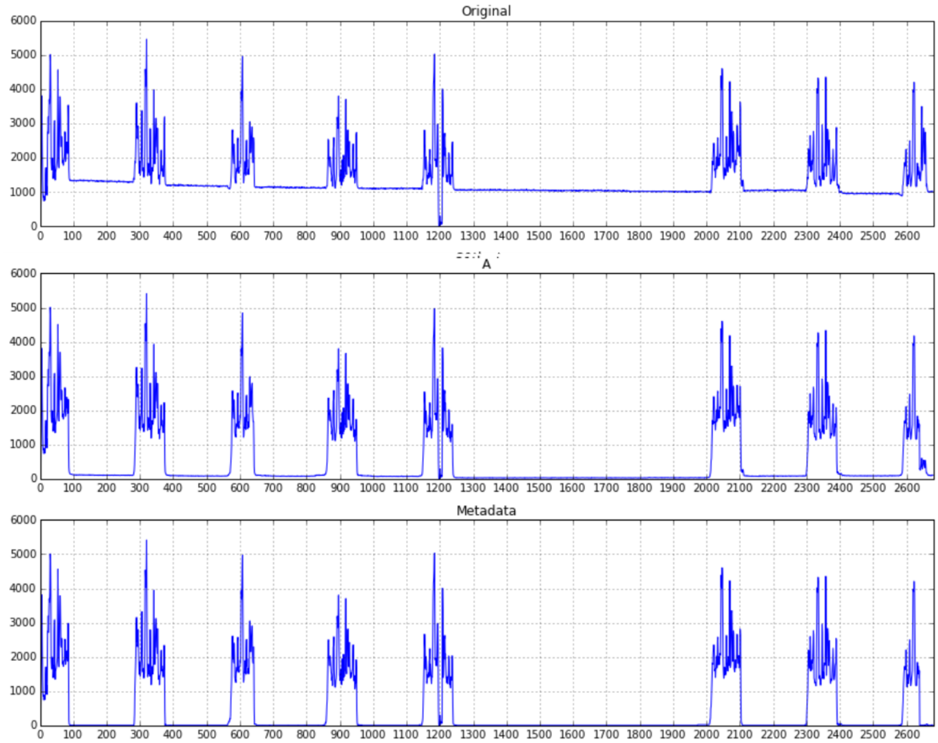

Fig. 1: **Total contact number measured in the original, metadata and cleaned (A) tensors with respect to time.** An obvious amount of contacts in the original state is distributed along the entire time-line. By contrast, the cleaning procedure managed to identify and erase most of the anomalies.

The results after 23 iterations of the iterative framework presented in Section 2 are summarised in Fig. 1, that shows the total contact number evolving in time measured in the original tensor and in the cleaned tensor generated by the iterative process. The total amount of contacts during the school closure is extremely reduced as a result of the cleaning process. Normal interactions belonging to the classes emerge and meaningful patterns are recovered.

In order to validate the method, a reference tensor was created and used as a ground truth. To this end, anomalous behaviours were identified and removed from the original dataset by applying the step detection on the temporal contact

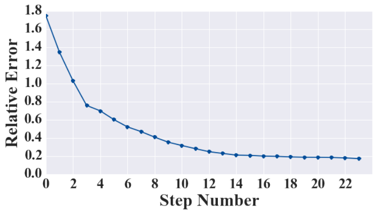

Fig. 2: Relative error, computed at each iteration by using the L_1-norm.

Table 1: Retrieval measures of tensor entry labelling. The entries classified as anomalous or meaningful are respectively marked as 0 or 1.

	Precision	Recall	F1-score
0	0.98	0.92	0.95
1	0.88	0.96	0.92

densities of each class. The ensuing tensor was compared with the cleaned tensor $\mathcal{T}'$ at each iteration, by computing the relative error with the L_1-norm. This quantity, shown in Fig. 2, monotonically decreases with the number of steps of the iterative approach and stabilizes around the low value of 0.2.

The reference tensor was then used to label the entries of the original tensor as anomalous or meaningful, in order to compute the confusion matrix summarizing the performances of the iterative approach. The resulting recall and precision values at each entry level reported in Tab. 1 and the overall accuracy of 0.94 highlight the high performance of the iterative approach.

4 Conclusions

Time-varying graphs can expose both topology and temporal correlations, which make the anomaly detection a major challenge. The iterative approach introduced here captures such correlations and enables to discriminate between meaningful and anomalous patterns. The evaluation measurements of the anomaly detection, achieved on the primary school dataset, highlights the high performance of the implemented method.

This iterative method is a principled approach, which provides an unsupervised way to identify and select meso-scale data anomalies. However, some limitations are worth noting. The method relies on the temporal activity profile of a latent component to be able to classify it as anomalous or not. Moreover, the implication of the NTF makes the iterative approach computationally costly, in terms of memory and time. The latter problem is being tackled as a lot of research is devoted to improve the efficiency of the implementation.

Finally, the iterative framework could be extended to the case of a greater number of dimensions, e.g. with sensors localized in space.

5 Acknowledgements

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.

References

1. W. Eberle, L. Holder, Discovering structural anomalies in graph-based data. In *ICDM Workshops*, pp. 393-398, 2007.
2. M. Mongiovi, et al., Netspot: Spotting significant anomalous regions on dynamic networks. *SIAM International Conference on Data Mining*, 2013.
3. P. Holme, J. Saramäki. Temporal networks. *Phys. Reps.* 519:97-125, 2012.
4. J. Stehlé, N. Voirin, A. Barrat,C. Cattuto, L. Isella, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS One* 6:e23176, 2011.
5. C. Cattuto, et al. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE* 5:e11596, 2010.
6. T.G. Kolda and B.W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455-500, 2009.
7. H. Kim, L. Eldé n, and H. Park. Non-negative tensor factorization based on alternating large-scale nonnegativity-constrained least squares. In *Proceedings of IEEE 7th International Conference on Bioinformatics and Bioengineering* (BIBE07), volume II, pages 1147-1151, 2007.
8. J. Kim and H. Park. Fast nonnegative tensor factorization with an active-set-like method. In *High-Performance Scientific Computing*, Springer, London, 2012, pp. 311-326
9. R. Bro and H. A. L. Kiers, A new efficient method for determining the number of components in PARAFAC models, *Journal of Chemometrics*, 17 (2003), pp. 274-286.
10. E.E. Papalexakis and C. Faloutsos, Fast efficient and scalable core consistency diagnostic for the parafac decomposition for big sparse tensors, in *Acoustics, Speech and Signal Processing* (ICASSP), 2015 IEEE International Conference on. IEEE, 2015.
11. M. Fang, G. Yue, and Q. Yu, The Study on An Application of Ostu Method in Canny Operator, *Proceeding of the 2009 International Symposium on Information Processing*, Huangshan, P. R. China, pp. 109-112, August 2009.