Temporal Pattern Mining from Evolving Networks

Angelo Impedovo

University of Bari "Aldo Moro", Department of Computer Science
Knowledge Discovery and Data Engineering Laboratory, Bari 70125, Italy,
angelo.impedovo@uniba.it

Abstract. The description of the evolution of a network is a known problem. Several data mining algorithms try to describe relevant aspects of static networks, the challenge of describing the temporal dynamics is quite recent. The study of the evolution of a network must consider the analysis of the changes exhibited over time. In this work we face the problem of characterization of the temporal dynamics of an evolving network using descriptive models of the changes exhibited over time.

Keywords: Evolving networks, Frequent Pattern Mining, Data Mining

1 Introduction

The study of evolving networks is an emerging discipline useful to many scientific fields. Available tools can be used anytime it is possible to study a phenomenon directly or indirectly convertible to a network-based representation. In fact, many complex systems from the real world can be easily described in terms of a network of interconnected components, this is the suitable representation for modeling complex objects [1], since it can naturally deal with the large variety of types of entities (nodes) and relationships (edges) among the entities. Therefore, the analysis of the properties of a real complex system can be translated into the analysis of the network used to describe it.

Some real complex systems may evolve over time, this evolution affects the networks used to describe them. The evolution of networks has been modeled by different mathematical models that describe it according to specific laws [2]. The usage of such mathematical models can limit the analysis, they characterize the evolutionary process considering global features of the network (degree, density, diameter, etc.) only, forgetting any information local to nodes, edges or to entire subnetworks. However, even if many networks may evolve accordingly to a common evolutionary schema only, many others networks do not evolve considering a single schema. Furthermore, many mathematical models are unrealistic and, therefore, are not suitable to model the evolution of real networks.

Existing data mining approaches assume that the evolution of a network is entirely determined by a stochastic process whose parameters may vary over time. Every detail about the process is unknown to the observer which would

like to infer them starting from observed data without making any prior assumption about the underlying evolutionary model. In this perspective what become relevant are the properties of nodes and edges that, in their turn, can also be described as complex objects with their own descriptive attributes. Some applications of the evolving network analysis may concern: the discovery of social events from social media, the evolution of citations in scientific collaboration networks and, in the end, the analysis of spreading diseases in contagion networks.

The work is part of a broad research project in his early stage. We tackle the extraction and the characterization of temporal dynamics of evolving networks through the learning of descriptive models expressed in form of sets of patterns, this is done without making any prior assumption about the nature of the underlying evolutionary models of the networks. The rest of this paper is organized as follows: related works are introduced and discussed in the section 2, the novelty and the contributions of the proposal is clarified in section 3, then conclusions are drawn in section 4.

2 Literature review

Numerous contributions can be gathered under the problem of mining dynamics in evolving networks. Two main approaches can be identified: a clustering-based approaches and a pattern-based approaches. Unsupervised techniques are more attractive and viable than supervised ones because of the absence of any ground truth able to establish which data represent a change and which do not.

Pattern-based approaches relies on the frequent pattern mining framework. In [3], the authors identify subgraphs changing over time by means of vertex-importance scores and vertex-closeness changes in subsequent snapshots of the network. Berlingerio et al. [4] proposed the extraction of graph evolution rules from a sequence of snapshots of an evolving graph by resorting to frequent pattern mining solutions. In [5, 6] the dynamics are characterized introducing the notion of evolution chains, changes discovered over consecutive time-periods and then combined in sequence to model the evolution of the network.

However, the clustering-based approaches focuses on the changes of network-based or node-based indicators, thus they lead to discovering changes that regards only the whole network or some nodes, without any information on the topology. The pattern-based techniques work on the topology, as they works considering subgraphs, and may provide insightful information as they operate on portions of the whole network.

3 Possible Approaches

Different pattern-based approaches to the problem of mining changes from evolving networks will be given in this research project. The approaches will allow to: i) discover the temporal collocation of changes, ii) characterize the changes through the learning of descriptive models starting from observed snapshots.

The temporal collocation of changes is found by an analysis based on time windows [7], in which network snapshots can be aggregated. Working on time-windows allows us to summarize the changes occurring at the level of time instants and model them at a higher level of temporal granularity, that is, intervals of time. More formally, a time window of size n is an ordered sequence of snapshots $W = \langle G_0, G_1, ..., G_n \rangle$ in which every snapshot G_i is associated to a specific discrete time point t_i. Two time windows W_2, W_2 are said to be consecutive if $W_1 = \langle G_0, ..., G_j \rangle$ e $W_2 = \langle G_{j+1}, ..., G_k \rangle$.

In order to build descriptive models that reflect the temporal collocation of changes it is necessary to spot the optimal size of the windows. In other words an optimal partitioning of the snapshots has to be found. An important contribution of the present work lies in the possibility to partitioning the snapshots both in a fixed and adaptive manner. The fixed partitioning is done by splitting data between consecutive time windows of equal size. The adaptive partitioning is done using automatic criteria able to find time points associated to significant variations of the distributions of observed edges and nodes.

The problem to build descriptive models of the temporal dynamics can be resolved using frequent patterns. The usage of frequent patterns allows to abstract the network data, with the advantage to not work directly on nodes and edges, which requires for more computation. Patterns summarize local portions of the network (subnetworks), whereas frequent patterns denote subnetworks conserved over the snapshots. The relative frequency of a pattern P with respect to a time window W is the fraction of snapshots in W of which P is a subnetwork: $freq(P, W) = \frac{|\{G_i \in W | P \subseteq G_i\}|}{|W|}$. Consequently a pattern is said to be frequent in W if his relative frequency in W exceeds a minimum user-defined threshold value α, $freq(P, W) > \alpha$.

In this work frequent patterns will be used as a way to denote features of the network that are stable over time. Therefore, it become possible to think about the changes in terms of variations of the frequent subnetworks spotted. Such variations may reflect changes of the network with a certain level of statistical confidence, also they may concern any change regarding both the statistical parameters and the topology of the network.

Fundamental contribution of this work lies in the characterization of 3 types of change: i) emerging changes, ii) trend-based changes, iii) periodical changes.

i) Emerging changes are variations of conserved subnetworks and occur when their frequency significantly vary between two consecutive time-windows. To discover emerging changes, we resort the notion of emerging patterns (EPs) [8]. Emerging changes quantify substantial changes in the frequencies of specific subnetworks between two consecutive windows W_1 and W_2 by means their growth-rate:

$$GR(P, W_1, W_2) = \frac{freq(P, W_1)}{freq(P, W_2)} \in [0, +\infty) \tag{1}$$

A subnetwork P is said to be emerging if the associated growth-rate exceeds a minimum user-defined threshold value β, more briefly if $GR(P, W_1, W_2) > \beta$.

ii) Differently from emerging changes, trend-based changes concern gradual variations of conserved subnetworks that cannot be spotted through a pairwise analysis of the windows, they manifests over a relatively longer sequence of windows in which the frequency of conserved subnetworks monotonically increase or decrease. To discover trend-based changes we evaluate the relative frequency monotonicity of a subnetwork. Given the sequence of m consecutive time windows $T = \langle W_1, W_2, ...W_m \rangle$, a subnetwork P, a sign $\psi \in \{+, -\}$ such that $\forall i \in \{1, ..., m-1\}$ then:

$$\begin{aligned} \psi = + &\iff freq(P, W_i) < freq(P, W_{i+1}) \iff GR(P, W_{i+1}, W_i) > 1 \\ \psi = - &\iff freq(P, W_i) > freq(P, W_{i+1}) \iff GR(P, W_i, W_{i+1}) > 1 \end{aligned} \tag{2}$$

The triple (P, T, ψ) is called trend-based change. If $|T| > 2$ the trend of increase/decrease is said to be global, spanning through a larger period of time, resulting more interesting. Trend discovery is a challenging task, in fact it is unlikely that real network whose frequency monotonically vary without any fluctuation over a sequence of time windows. The problem will be tackled comparing the relative frequency $freq(P, W_m)$ with an aggregate value, computed over all previous windows, such as the average relative frequency λ:

$$\lambda = \frac{1}{m-1} \sum_{i=1}^{m-1} freq(P, W_i), \quad \psi = \begin{cases} + \iff freq(P, W_m) > \lambda \\ - \iff freq(P, W_m) < \lambda \end{cases} \tag{3}$$

iii) Periodic changes reveal regularly occurring changes over time. The periodicity is a good indicator for the relevance of a change, this may refer to events regularly occurring and then more interesting than those episodic. Another way to search for periodic changes is to study the growth-rate periodicity [9] of a subnetwork. Given the sequence $T = \langle W_1, W_2, ...W_m \rangle$ of m consecutive time windows, the subnetwork P is said to be periodic of period $\pi > 0$ if emerges with the same growth-rate every π windows:

$$\begin{gathered} GR(P, W_i, W_{i+1}) = GR(P, W_{i+k\pi}, W_{i+k\pi+1}) \\ |i - (i + k\pi)| = k\pi,\ k \in \mathbb{N} \end{gathered} \tag{4}$$

Periodic change discovery is a challenging task, it is unlikely that real networks exhibit specific subnetworks in a exactly periodic way. Main difficulties are: i) spotting exactly periodic changes of period π, ii) detecting changes that are periodically quantified by the same numerical value of growth-rate. The limitations will be overcome by relaxing the given definition in order to characterize nearly-periodic subnetworks. Given an arbitrary tolerance level $J > 0$, the set Ψ of categorical values and a mapping $\Theta : [0, +\infty) \to \Psi$, then:

$$\begin{gathered} \Theta(GR(P, W_i, W_{i+1})) = \Theta(GR(P, W_{i+k\pi}, W_{i+k\pi+1})) \\ k\pi - J \leq |i - (i + k\pi)| \leq k\pi + J,\ k \in \mathbb{N} \end{gathered} \tag{5}$$

instead of the equality test based on raw numerical values of the growth-rate, an equality test between categorical values is preferred.

4 Final remarks and ongoing works

Evolving networks are powerful tools used to describe the evolution of real-world complex systems, their analysis may help to better comprehend the temporal dynamics and the evolution of such complex systems. The pattern-based representation, together with considerations about the monotonicity and periodicity of the growth-rate, may help to better characterize the relevance of the changes. The application of temporal pattern mining approaches to the analysis of evolving networks deserves more investigation. We are currently performing experiments to evaluate the effectiveness of the emerging changes discovery as well as a distributed implementation of the proposed methods, in order to be able to process large-scale networks.

Ideas and contributions proposed in this work are part of the doctorate program of the author, which is supervised by prof. Michelangelo Ceci and dr. Corrado Loglisci.

References

1. Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012.
2. Deepayan Chakrabarti and Christos Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM computing surveys (CSUR)*, 38(1):2, 2006.
3. Zheng Liu, Jeffrey Xu Yu, Yiping Ke, Xuemin Lin, and Lei Chen. Spotting significant changing subgraphs in evolving graphs. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 917–922. IEEE, 2008.
4. Michele Berlingerio, Francesco Bonchi, Björn Bringmann, and Aristides Gionis. Mining graph evolution rules. *Machine learning and knowledge discovery in databases*, pages 115–130, 2009.
5. Corrado Loglisci, Michelangelo Ceci, and Donato Malerba. Discovering evolution chains in dynamic networks. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 185–199. Springer, 2012.
6. Corrado Loglisci, Michelangelo Ceci, and Donato Malerba. Relational mining for discovering changes in evolving networks. *Neurocomputing*, 150:265–288, 2015.
7. Joao Gama and Mohamed Medhat Gaber. Learning from data streams: Processing techniques in sensor networks. *ISBN 3540736786, 9783540736783*, pages 31–33, 2007.
8. Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52. ACM, 1999.
9. Corrado Loglisci, Michelangelo Ceci, Angelo Impedovo, and Donato Malerba. Mining spatio-temporal patterns of periodic changes in climate data. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 198–212. Springer, 2016.