Mining Microscopic and Macroscopic Changes in Network Data Streams (Discussion Paper)

Corrado Loglisci, Angelo Impedovo, Michelangelo Ceci, Donato Malerba

Dept. of Computer Science, University of Bari "Aldo Moro", Bari, Italy
{corrado.loglisci, angelo.impedovo, michelangelo.ceci, donato.malerba}@uniba.it

Abstract. Network data streams offer an abstraction of complex systems from the real-world, which can be seen as producers of unbounded sequences of complex data generated at high speed. Many complex systems evolve according to stochastic processes which remain unknown to the interested users. As a consequence, changes happen in an unpredictable manner and may involve various portions of the observed complex systems. In this scenario, an interesting problem concerns the identification and characterization of the changes that may concern both the whole structure of a complex system and small parts of it. We conjecture that the former can be explained by the latter and conversely, the latter can trigger the former. This type of problem requires a quite holistic strategy that traditional approaches often do not carry out because they focus on either the whole network or on some portions only. In this discussion paper, we describe a descriptive data mining approach based on frequent pattern discovery that we designed for recent research work. It combines frequent pattern with automatic time-window setting, in order to identify and characterize *macroscopic changes* and *microscopic changes* as changes that have an impact on a substantial part of the network or on specific portions, respectively. We provide arguments of the viability to real-world applications through two case studies, more precisely, telecommunication networks and geo-sensor networks.

Keywords: Change Mining, Network Streams, Frequent Subnetworks

1 Introduction

The dissemination of technologies able to unceasingly record information from real-world applications has lead to the development of systems based on data stream models. Often such systems work in highly dynamic and complex domains where data are naturally interconnected. This is the case of social networks and telecommunication infrastructures. Analyzing data continuously generated

in these fields requires data stream mining algorithms that operate on network data. In this scenario, a natural and interesting problem is that of analyzing data streams in order to identify the changes in the network interactions over time. However, in the domains in which human supervision or reference knowledge are not promptly available we cannot learn models able to recognize changing behaviors, and therefore the supervised approaches give way to those unsupervised. The descriptive data mining techniques and, in particular, the pattern-based approaches are gaining great attention thanks to their peculiarities to provide arguments to the change-related discoveries through statistical evidence. In [1], the discovery of *graph evolution rules* starting from sequence of graph snapshots from an evolving graph has been proposed. The rules are graphs in which nodes denote entities while edges are labeled with the time-stamp. However, this approach is only able to represent insertions and not deletions of nodes/edges. In [6], we studied how characterizing the evolution of specific portions of the whole network by introducing the notion of evolution chains. In [9], the authors track changes in the trend of statistical parameters of frequent patterns in social networks. The patterns are mined from descriptive properties of the nodes of the networks and therefore do not provide information on the structural changes.

We addressed the critical points of the aforementioned works through a method, described in detail in [5], which contributes to the problem of change mining in three different perspectives, that is, the task studied, methodological approach and algorithmic solution. As for the task, we argue that changes concerning the global properties of a network can be ascribed to a combination of local variations. Changes of single portions (microscopic changes) can trigger changes in the global properties of the network (macroscopic). More specifically, microscopic changes are associated with subnetworks whose properties change over time, whereas macroscopic changes are associated with composite aggregations of microscopic changes.

As for the methodological approach, we resort to the frequent pattern mining framework to generate a summarizing form of the network data on which identifying evolutions. Therefore, frequent subnetworks are not the primary objective of the paper, but the means to capture two kinds of changes. By using pattern-subnetworks, we can search for changes in an abstract description of the network [4], while alleviating the computational cost that the analysis of raw data would require. So, single frequent subnetworks synthesize specific portions, while sets of frequent subnetworks synthesize larger portions of the network. This methodological decision is motivated by the fact that frequent subnetworks provide arguments for the robustness of our method. Indeed, frequency denotes statistical evidence, therefore, frequently occurring changes turn out to be more interesting than episodic ones, because they are replicated and, probably, well-established over time.

As for the algorithmic solution, the hypothesis on the absence of labeling of the changes suggest us to monitor data streams and decide when new data snapshots exhibit a change, that is, when the network differs from the past. To this end, we use two time-windows, one to collect data snapshots of the "re-

cent past" and the other to collect new data snapshots. This way, we capture the status of the network in two distinct time-windows, keep patterns and their statistical properties updated, and adapt the search strategy to the variations of the underlying data distribution. The combined use of time-windows and frequent subnetworks suggests us to search for *i)* macroscopic changes as variations between sets of frequent subnetworks discovered on two time-windows, and *ii)* microscopic changes as punctual variations of frequent subnetworks occurring at the level of data snapshots. The use of two windows ("recent past" and "new snaphosts") has been studied also for predictive tasks [8].

2 Basics and Problem Statement

In the following, we provide some basic concepts as well as the notions of change of interest for this work. A network data stream is the time-ordered sequence $D = \langle G_1, G_2, \ldots \rangle$ of network snapshots G_i observed at time-point t_i. Each snapshot G_i is labeled graph with labeled edges $G_i \subseteq N \times N \times L$, where N is the set of nodes and L is set of edge labels. In particular a *landmark window* $W = [t_i, t_j]$ of width $|W| = j - i + 1$ is the sequence of consecutive snapshots $\{G_i, \ldots, G_j\}$. Following [3], a (landmark) window W' is successive to another (landmark) window W when they share some initial time-points, that is, $W = [t_i, t_n]$ and $W' = [t_i, t_m]$, with $t_i < t_n < t_m$.

The proposed approach relies on the *frequent pattern mining* framework. In particular, we i) restrict the pattern language to *subnetworks* in which there exists a path between any two nodes, and ii) mine *frequent subnetworks* from snapshots of a window W. The *support* of a subnetwork S in W is defined as $sup(S, W) = \frac{|\{G_i \in W | S \subseteq G_i\}|}{|W|}$. A subnetwork S is *frequent* in W if $sup(P, W) \geq minSUP$, where $minSUP \in [0, 1]$. Then, F_W denotes the set of all the frequent subnetworks discovered from the window W. We consider two types of changes:

Definition 1 (Macroscopic change). *Given* $minMC \in [0, 1]$*, a macroscopic change is found if* $MC(W, W') = \frac{|F_{W'} - F_W| + |F_W - F_{W'}|}{|F_W| + |F_{W'} - F_W|} > minMC$.

Definition 2 (Microscopic change). *Given* $minGR \in [1, +\infty)$*, a subnetwork* S *denotes a microscopic change when i)* $\frac{sup(S,W)}{sup(S,W')} \geq minGR$ *if* $S \in (F_W - F_{W'})$*, or ii)* $\frac{sup(S,W')}{sup(S,W)} \geq minGR$ *if* $S \in (F_{W'} - F_W)$.

where, $W = [t_i, t_n]$ and $W' = [t_i, t_m]$ are two successive windows, F_W and $F_{W'}$ are the sets of frequent subnetworks mined on W and W', respectively.

In these terms, the problem of identifying and characterizing macroscopic and microscopic changes in a network data stream $D = \langle G_1, G_2, \ldots \rangle$ can be interpreted as the search for pairs of successive windows (W, W'), in which the *i)* quantification of the variations between F_W and $F_{W'}$ exceeds the threshold $minMC$ and *ii)* quantification of the variations the support of S (either $S \in (F_W - F_{W'})$ or $S \in (F_W - F_{W'})$) exceeds the threshold $minGR$.

3 The proposed mining approach

In this discussion paper, we discuss the KARMA algorithm proposed in [5], which delivers a computational solution to discover the changes before formulated. The algorithm (Figure 1) iteratively consumes blocks Π of network snapshots coming from the stream D (Step 2) by using two successive landmark windows W and W' (Step 3). This way, it mines frequent subnetworks, F_W and $F_{W'}$, necessary to the identification of both macroscopic and microscopic changes (Steps 4-5). The window grows ($W = W'$, Step 8) with new network snapshots, and the associated set of frequent subnetworks is kept updated until $MC(W, W') > minMC$ and a macroscopic change is found. In that case, the algorithm mines the microscopic changes (Step 6) and drops the content of the window by retaining only the last block of transactions ($W = \Pi$, Step 7). Then, the analysis restarts.

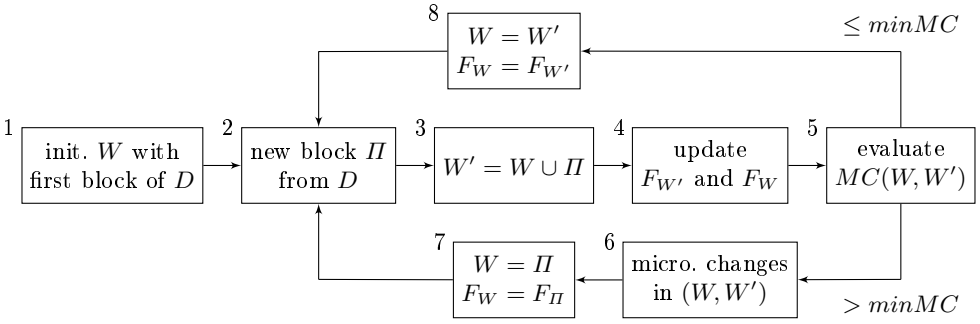

Fig. 1. The KARMA algorithm flowchart

The algorithm relies on the frequent subgraph mining framework, which is intractable for massive datasets due to the combinatorial explosion of the frequent subgraphs. For this reason, the sets F_W and $F_{W'}$ are kept updated and not recomputed from scratch upon the arrival of new transactions. Algorithmically, this is done by means of incremental solutions of frequent pattern mining.

3.1 Macroscopic change mining

When a new block of network snapshots Π is acquired, KARMA builds the window W' ($W' = W \cup \Pi$) and checks if there is a *macroscopic change* by matching W' against W. To do that, KARMA offers two peculiarities: i) identification of relevant changes on a summary of the data (the set of frequent subnetworks) rather than on the raw data, and ii) quantification of the changes in terms of the differences between the set F_W of frequent subnetworks discovered on W and the set $F_{W'}$ of frequent subnetworks discovered on W'. It is desirable to seek for changes on data by looking at frequent subnetworks in order to avoid false alarms, thus achieving higher robustness to noisy data. In this scenario, a

single noisy network snapshot (e.g. a noisy outlier) would not affect the set of frequent subnetworks, depending on the minimum support $minSUP$. In fact, the main assumption behind the *macroscopic change* is that the whole network does not exhibit a change between W and W' if the frequent subnetworks, F_W and $F_{W'}$, do not significantly change. To measure the amount of change, KARMA computes the macroscopic change (MC) as $MC(W, W') = \frac{|F_{W'} - F_W| + |F_W - F_{W'}|}{|F_W| + |F_{W'} - F_W|}$. Where W and W' are two successive landmark windows, $|F_W - F_{W'}|$ denotes the number of subnetworks which are frequent in W and infrequent in W', and $|F_{W'} - F_W|$ is the number of subnetworks which are frequent in W' and infrequent in W. The formula quantifies the fraction of subnetworks which have crossed the minimum support threshold $minSUP$, that is, those which were frequent (infrequent) in W and become infrequent (frequent) in W', and which therefore indicate a relevant change in the underlying network data distribution.

3.2 Microscopic change mining

In KARMA, the discovery of *microscopic changes* is performed only when a *macroscopic change* is spotted. Indeed, a microscopic change accounts for the contribution that each subnetwork, which crosses the minimum support threshold $minSUP$, gives to the macroscopic change. To do that, we resort the notion of *emerging pattern* [2], which have been proven useful in the classification setting when mining discriminative patterns between two classes. In KARMA, we mine *emerging subnetworks* whose support significantly spreads between W and W'. This is done by quantifying the *growth-rate* of the subnetwork S, $\frac{sup(S,W)}{sup(S,W')} \geq minGR \quad if \quad S \in (F_W - F_{W'})$ (or alternatively, $\frac{sup(S,W')}{sup(S,W)} \geq minGR \quad if \quad S \in (F_{W'} - F_W)$). Every emerging subnetwork S satisfying the growth-rate inequality denotes a single *microscopic change*. The $minGR$ threshold is a tradeoff between the completeness and the simplicity of the descriptive model. Low values of minGR lead to expressive models, while high values lead to synthetic models about the change.

3.3 KARMA at work

We show the applicability of KARMA for the analysis of a real-world network by discussing some discoveries and commenting on their actionability with respect to facts and events occurred in the domain. In the following, we discuss two case studies: the first39 in the domain of telecommunication networks (NODOBO dataset), and the second in the domain of geo-sensor networks (NOAA dataset).

The *NODOBO*[1] dataset concerns a communication network and contains telecommunication transactions gathered during a study of the mobile phone usage of 27 students of a Scottish state high-school, from September 2010 to February 2011. The students communicate by phone calls, text messages (SMS) and Bluetooth connectivity. When building the network, the nodes represent the students, while the edges represent the different modalities of communication.

[1] http://nodobo.com/release.html

The dataset $NOAA$[2] was developed in the context of the Reanalysis project by the National Center for Environmental Prediction and the National Center for Atmospheric Research. The project aimed at providing new atmospheric analysis by gathering daily measurements of various meteorological quantities (e.g. relative humidity and air temperature) by means of geo-localized sensors equally distributed over space. In this work, we built the network with the measurements of relative humidity of the time-interval January, 1st 1990 - December, 31st 2010, recorded daily on an area that roughly covers North-Central America. The nodes of the network represent the sensors, while the edges are nominal values denoting the relative humidity values measured on the two linked sensors.

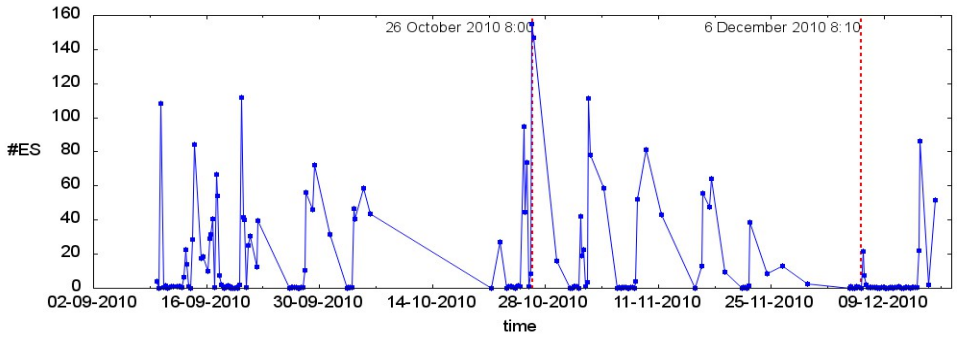

Fig. 2. Number of microscopic changes ($\#ES$) discovered from the NODOBO dataset as a function of time. The values of the $\#ES$ have been multiplied by 10^{-3}.

As for the analysis of NODOBO, in Figure 2, we see a succession of points with a decreasing trend, which has the peak on October 26th, 2010 and the lowest number of microscopic changes on December 6th, 2010. To give a practical interpretation to this behavior, it is useful to say that in Scottish state high-schools there is a holiday period which covers the second and third week of October, thereafter the school activities continue. So, the projection of Figure 2 reveals that, when the school activities resume, there is high variability (many microscopic changes) in the modalities of communication, which, as time goes by, tends to decrease. This may provide indications on the use of mobile phone of the students, which can be exploited, for instance, to plan the school policies and to improve the mobile network services in the area.

Among the microscopic changes associated to the peak, we find the pattern P={(*student_14, student_0, bluetooth*), (*student_18, student_0, bluetooth*), (*student_2, student_0, high_length*)}. It is involved in the strongest macroscopic change (quantified as 0.94), which starts on October 26th 2010 at 3:00 (the time-point after the window [2010 Oct 25-22:00, 2010 Oct 26-2:55]) and terminates on October 26th 2010 at 7:55. Specifically, P denotes the doubling

[2] https://coastwatch.pfeg.noaa.gov/erddap/griddap/esrlNcepRe.html

(growth-rate equals to 2.0) of the occurrences of the associated subnetwork from the window [2010 Oct 25-22:00, 2010 Oct 26-2:55] to the window [2010 Oct 25-22:00, 2010 Oct 26-7:55]. On the contrary, (we verified) P does not appear in the set of microscopic changes corresponding to the successive macroscopic change detected between the landmark windows [2010 Dec 05-22:10, 2010 Dec 06-3:05] and [2010 Dec 05-22:10, 2010 Dec 06-8:05], which may indicate that the modalities of interaction among the three students become stable.

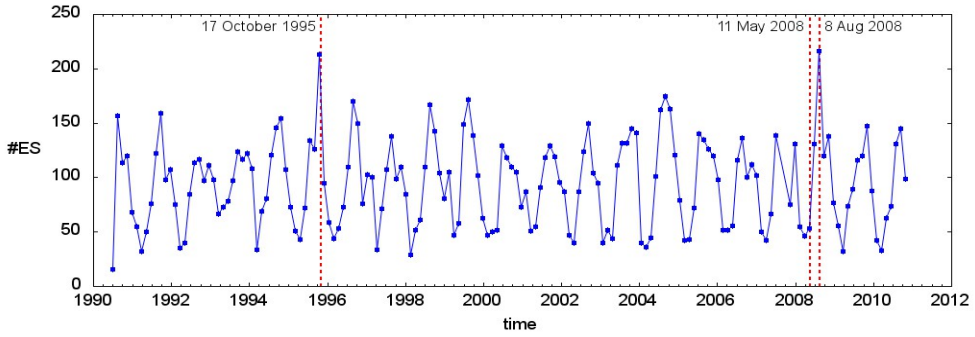

Fig. 3. Number of microscopic changes ($\#ES$) discovered from the NOAA dataset as a function of time.

As for the NOAA domain (Figure 3), there are several macroscopic changes, those which have a greater impact on the network are characterized by more than 150 microscopic changes. In particular, there are two macroscopic changes with the highest number, they start on October 17th, 1995 and August 2008, 9th respectively. We deepened the microscopic changes corresponding to the two points and spotted several emerging subnetworks in common, one is P={(*<10,300>, <10,305>, from_70_to_80*), (*<10,300>, <7.5,297.5>, from_80_to_90*), (*<12.5,297.5>, <7.5,297.5>, from_80_to_90*)}.

By mapping the nodes of P into a geodesic space, we see they identify two regions, both cover approximately the area of the state of Venezuela and part of the Caribbean sea, where there are small differences in terms of relative humidity (the edge labels refer to consecutive ranges). This meteorological scenario becomes less frequent (the growth-rate decreases) over the window [1995 Oct 17, 1995 Nov 30] and [2008 Aug 09, 2008 Sep 22] respectively, which suggests the possibility of different behavior, in the same geographic area, occurred before or after those two macroscopic changes.

4 Conclusions

In this discussion paper, we have investigated the problem of identifying relevant changes in network data streams, where the changes can be distinguished in two

categories: macroscopic changes and microscopic changes. The system presented in the paper, called KARMA, is able to simultaneously extract macroscopic changes and microscopic changes by exploiting the fact that they are inevitably related to each other. Two case studies have shown the usefulness and the actionability of the changes in the domain of geo-sensor networks and telecommunication networks. An extensive discussion on the influence of the parameters (minSUP, minMC, minGR) on the results can be found in [5], where a comparative evaluation of the running times is also reported. The interested reader can refer to the journal paper for further details.

For future work, we plan to investigate two main research directions: *i)* use of solutions of big data analytics to detect changes in very large networks, and *ii)* study of the closed patterns [7] to discover non-redundant subnetworks.

References

1. Berlingerio, M., Bonchi, F., Bringmann, B., Gionis, A.: Mining graph evolution rules. In: Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I. pp. 115–130. ECML PKDD '09, Springer-Verlag, Berlin, Heidelberg (2009)
2. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 15-18, 1999. pp. 43–52 (1999)
3. Gama, J., Gaber, M.M.: Learning from data streams: processing techniques in sensor networks. Springer (2007)
4. Loglisci, C., Berardi, M.: Segmentation of evolving complex data and generation of models. In: Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China. pp. 269–273. IEEE Computer Society (2006). https://doi.org/10.1109/ICDMW.2006.144
5. Loglisci, C., Ceci, M., Impedovo, A., Malerba, D.: Mining microscopic and macroscopic changes in network data streams. Knowl.-Based Syst. **161**, 294–312 (2018)
6. Loglisci, C., Ceci, M., Malerba, D.: Relational mining for discovering changes in evolving networks. Neurocomputing **150, Part A**(0), 265 – 288 (2015)
7. Loglisci, C., Malerba, D.: Mining multiple level non-redundant association rules through two-fold pruning of redundancies. In: Perner, P. (ed.) Machine Learning and Data Mining in Pattern Recognition, 6th International Conference, MLDM 2009, Leipzig, Germany, July 23-25, 2009. Proceedings. Lecture Notes in Computer Science, vol. 5632, pp. 251–265. Springer (2009). https://doi.org/10.1007/978-3-642-03070-3_19
8. Loglisci, C., Malerba, D.: Leveraging temporal autocorrelation of historical data for improving accuracy in network regression. Statistical Analysis and Data Mining **10**(1), 40–53 (2017). https://doi.org/10.1002/sam.11336, https://doi.org/10.1002/sam.11336
9. Nohuddin, P.N.E., Coenen, F., Christley, R., Setzkorn, C., Patel, Y., Williams, S.: Finding "interesting" trends in social networks using frequent pattern mining and self organizing maps. Knowl.-Based Syst. **29**, 104–113 (2012)