A Method for Characterizing Communities in Dynamic Attributed Complex Networks

Günce Keziban Orman[#*], Vincent Labatut[*], Marc Plantevit[+], Jean-François Boulicaut[#]

[#] *INSA-Lyon, CNRS, LIRIS, Université de Lyon*
UMR5205, F-69621, Lyon, France
[4]jean-francois.boulicaut@insa-lyon.fr

[*]*Computer Engineering Department, Galatasaray University*
Ortaköy 34349, Istanbul, Turkey
[1]korman@gsu.edu.tr, [2]vlabatut@gsu.edu.tr

[+]*Université Lyon 1, CNRS, LIRIS, Université de Lyon*
UMR5205, F-69622, Lyon, France
[3]marc.plantevit@liris.cnrs.fr

Abstract—**Many methods have been proposed to detect communities in complex networks, but very little work has been done regarding their interpretation. In this work, we propose an efficient method to tackle this problem. We first define a sequence-based representation of networks, combining temporal information, topological measures and nodal attributes. We then describe how to identify the most emerging sequential patterns of this dataset and use them to characterize the communities. We also show how to highlight outliers. Finally, as an illustration, we apply our method to a network of scientific collaborations.**

***Keywords*: Dynamic Attributed Networks, Community Interpretation, Topological Measures**

I. Introduction

Complex networks have become very popular as a modeling tool, because they help to better understand the intrinsic laws and dynamics of complex systems in many fields: sociology, physics, genetics, computer, etc. [1]. The complex nature of the modeled systems leads to the presence of non-trivial topological properties in the corresponding networks. Among them, the *community structure* is one of the most common and most studied. Intuitively, we can define a *community* as a group of nodes which are densely interconnected, relatively to the rest of the network [2]. Hundreds of algorithms have been proposed to detect community structures [3].

Although these algorithms differ in terms of nature of the detected communities, algorithmic complexity, result quality and other aspects, their output can always be basically described as a list of node groups. From an applicative point of view, the question is then to make sense of these groups relatively to the studied system. For the community structure to be useful, it is necessary to interpret the detected communities. And yet, almost all works in the field of community detection concern the definition of detection tools. Only a very few works try to tackle the problem of characterizing and interpreting the communities.

Authors historically interpreted their data manually [4] but this somewhat subjective approach does not scale well on large networks. In order to characterize each community individually, some authors take advantage of the information conveyed by nodal attributes, when it is available. In [5], the authors propose a statistical method to characterize the communities in terms of the over-expressed attributes found in the elements of the community. In [6], authors interpret the communities of a social attributed network by using statistical regression and discriminant correspondence analysis. Community detection methods based on nodal attributes also allow outputting the most representative attributes of the communities [7, 8]. These works are valuable, however they do not take advantage of all the available information to enhance the interpretation process.

In this work, we see the interpretation problem as independent from the approach used to detect the communities, and we try to propose a method tackling the limitations of the existing approaches. We use descriptors to represent the nodes, considering both their attributes and topological properties. We then represent a dynamic attributed network as a sequence database of node descriptors. We aim at finding the most representative emerging sequential patterns for each community on this new representation of the network. These patterns can then be used for both characterizing the community, and identify outliers. We illustrate our proposal on a dynamic co-authorship network extracted from DBLP. Our work contributes to the domain in several ways: (1) statement of community characterization as a specific problem, distinct from community detection; (2) introduction of a new representation of dynamic attributed networks, under the form of a sequence database; (3) definition a method taking advantage of this representation to extract sequential patterns able to characterize the communities, (4) illustration of our method through its application to a real-world network.

The rest of this article is organized as follows. In section II, we give the preliminary definitions needed to describe our method. In section III, we specify the problem and explain in detail our interpretation method. In section IV, we present our experimental results obtained on the DBLP data. Section V discusses our work and presents its possible extensions.

ASONAM 2014, August 17-20, 2014, Beijing, China
978-1-4799-5877-1/14/$31.00 ©2014 IEEE

II. Preliminary Definitions

We define a *dynamic network* $G = \langle G_1, \dots, G_\theta \rangle$ as a sequence of chronologically ordered time slices. Each *time slice* corresponds to a separated subnetwork G_t $(1 \leq t \leq \theta)$, which represents the connections between the nodes for a given time interval. Moreover, the networks we consider are *attributed*, meaning their nodes are described by some individual attributes. We note $n = |V|$ the number of nodes, which is constant through time. A *topological measure* quantifies the structural properties of the network or its components. Here, we focus on five nodal measures: *internal degree* (d), *local transitivity* (T), *within module degree* (z), *participation coefficient* (P) and *embeddedness* (e) [9]. The *within module degree* and *participation coefficien*t are two measures proposed by Amaral & Guimerà [10] to characterize the community role of nodes.

A *node descriptor* is either any of these five topological measures, or a node attribute. Let $D = \{D_1, D_2, \dots, D_k\}$ be the set of all descriptors. Each descriptor from D can take one of several discrete values, defined in its *domain* $\mathfrak{D}_i$ $(1 \leq i \leq k)$. We process each of these measures for each node, and at each time slice. A *community structure* $C = \{C_1, \dots, C_\lambda\}$ is a partition of V, where each $C_c \subseteq V$ and $\bigcap_{i,j \in \{1,\dots,\lambda\}} (C_i, C_j) = \emptyset$. The *community size* of a given community C_c is $|C_c|$, i.e. the number of nodes it contains. We note λ the number of communities in the community structure.

III. Characterization Method

We break down the problem of *community interpretation* in two sub-problems: 1) finding an appropriate way of *representing* a community, and 2) taking advantage of this representation to identify its most *characteristic features.* We solve the first sub-problem by representing a community as a set of sequences describing the evolution of its nodes. This encoding allows handling attributed dynamic networks, via their nodes topological measures and attributes. To solve the second sub-problem, we mine this set to identify sequential patterns fitting several criteria.

The process we propose includes 3 steps. The first is to identify the community structure. In the second step, we search for emerging sequential patterns and extract the corresponding supporting nodes for each community. Finally, the third step consists in choosing the most representative patterns to characterize the communities.

Step 1: Detecting Communities. To detect how nodes evolve in terms of community membership, we need first a reference community structure. To identify it, we chose to apply the so-called *sum method* [11]. First, we integrated the network links over time. Second, we applied the Louvain [4] algorithm, which is very widespread, to the resulting network.

Step 2: Mining Emerging Sequences. We want to characterize each community according to the common evolution of the descriptors of its nodes over time. For this purpose, we need to identify series of descriptor values which appear often in the same community and over several time slices. This is precisely the goal of sequential pattern mining methods.

Let us first define the concepts necessary to the description of the method itself. An *item* $(D_i, x) \in D \times \mathfrak{D}_i$ is a couple constituted of a descriptor D_i and a value x from its domain $\mathfrak{D}_i$. The set of all items is noted I. An *itemset* h is any subset of I. A *sequence* $s = \langle h_1, \dots, h_m \rangle$ is a chronologically ordered list of itemsets. The *size* m of sequence s is the number of itemsets it contains. A sequence $\alpha = \langle a_1, \dots, a_\mu \rangle$ is a *sub-sequence* of another sequence $\beta = \langle b_1, \dots, b_\nu \rangle$ iff $\exists i_1, i_2, \dots, i_\mu$ such that $1 \leq i_1 < i_2 < \cdots < i_\mu \leq \nu$ and $a_1 \subseteq b_{i1}, a_2 \subseteq b_{i2}, \dots, a_\mu \subseteq b_{i\mu}$. This is noted $\alpha \sqsubseteq \beta$. It is also said that β is a *super-sequence* of α, which is noted $\beta \sqsupseteq \alpha$.

The *node sequence* of a node v is a specific type of sequence noted $u(v) = \langle (l_{11}, \dots, l_{k1}) \dots (l_{1\theta}, \dots, l_{k\theta}) \rangle$ where l_{it} is the item containing the value of descriptor D_i for v at time t. A node sequence $u(v)$ includes θ itemsets, i.e. it represents all time slices. Each one of these itemsets contains all k descriptor values for the considered node at the considered time. In other words, $u(v)$ contains all the available descriptor-related data for node v. These tuples will be used later to constitute the database analyzed by our method.

The set of *supporting nodes* $S(s)$ of a sequence s is defined as $S(s) = \{v \in V : u(v) \sqsupseteq s\}$. The *support* of a sequence s, $Sup(s) = |S(s)|/n$, is the proportion of nodes whose node sequences are super-sequence of s. Similarly, the set of *supporting nodes* $S(s, C_c)$ of a sequence s in C_c is defined as $S(s, C_c) = \{v \in C_c : u(v) \sqsupseteq s\}$ and the *support* of a sequence in a community C_c, $Sup(s, C_c) = |S(s, C_c)|/|C_c|$, is the proportion of nodes in C_c, whose node sequences are super-sequence of s. Given a minimum support threshold noted min_{sup}, a *frequent sequential pattern* (FS) is a sequence whose support is greater or equal to min_{sup}. A *closed frequent sequential pattern* (CFS) is a FS which has no super-sequence possessing the same support.

In this study, we used the algorithm *CloSpan* [12] to identify all possible CFS for a given min_{sup}. CloSpan is an efficient algorithm, which can mine long sequences in practical time for real-world data. It outputs both the sequences and their supports. We modified the original CloSpan in a way to extract the supporting node sets $S(s, C_c)$. In our case, we want to identify, for each community, its most representative sequential pattern(s). For this purpose, we turn to the notion of *emerging pattern*, i.e. a pattern more frequent in a part of the node set than in the rest of it. The emergence of a pattern s relatively to a community C_c is measured by its *growth rate* $Gr(s, C_c) = Sup(s, C_c)/Sup(s, \overline{C_c})$ where $\overline{C_c}$ represents the nodes not belonging to C_c. A value larger than 1 means s is particularly frequent (i.e. emerging) in C_c, when compared to the rest of the network. We consider that the higher the growth rate, and the more representative the sequence s for community C_c. In order to calculate the growth rate, it would be necessary to search CFS in all communities separately, which can be a costly operation. However, we apply a more efficient method proposed in [13] to handle the case where classes are assigned to item sequences. For our problem, classes correspond to the communities.

Step 3: Selecting Sequential Patterns and Identifying Outliers. After the emerging patterns are identified for a given community, together with their support, growth rate and supporting nodes, we need to select the most *representative* ones, in order to characterize the considered community. We give more attention to the most emerging pattern, i.e. that whose growth rate is the highest. However, there is no guarantee for this pattern to cover a sufficient part of the community. And indeed, in practice it appears to be the opposite. It is thus needed to identify other complementary patterns, allowing us to obtain a more complete coverage of the community. Intuitively, we want to find a small number of patterns, such that they cover a significant part of the community, and are different in terms of supporting nodes. Thus, (1) the intersection of the patterns supporting nodes sets must be minimal, (2) the union of these supporting nodes sets must be maximal (if possible: the whole community), (3) the number of patterns must be minimal. In order to solve our problem, we select iteratively the most distant patterns, in terms of supporting node set. We use Jaccard's coefficient as a distance measure between the node sets.

Once the most characteristic patterns of a community have been identified (the most emerging one with its supplementary patterns, and the one with highest support), it is possible to use them to detect outliers. We determine the *outlier nodes set* as the set of nodes not following any representative pattern. These nodes are different in the sense they do not follow the general trends of their communities. We detect them automatically when finding the representative patterns.

The overall complexity of our method includes calculating all the selected topological measures, detecting the communities, applying CloSpan to identify the patterns, processing their growth rates, and finally selecting the most representative ones. Considering all simplifications and negligible terms, we get a worst case total complexity of $O(\lambda r^2)$ for the post-processing of calculating growth rate, where r is the number of patterns found by CloSpan.

IV. Results

We now present the results obtained on real-world data. We selected the dynamic co-authorship network extracted from the DBLP database. Each one of the 2145 nodes represents an author. Two nodes are connected if the corresponding authors published an article together. Each time slice corresponds to a period of five years. There are a total of 10 time slices, ranging from 1990 to 2012. The consecutive periods have a three year overlap for the sake of stability. For each author, at each time slice, the database provides the number of publications in 43 conferences and journals. We use this information to define 43 corresponding node attributes, and we add two more: the total number of conference and journal publications. Finally we have a total of 45 attributes.

For Guimerà &Amaral's measures, we use the thresholds originally defined in [10]. They distinguish community *hubs* ($z > 2.5$) and community *non-hubs* ($z \leq 2.5$). Note that this is different from the traditional notion of hub. A *community* hub is a node whose internal degree is well above its community average. The other topological measures are discretized depending on their distribution. The result community structure has 127 communities and a modularity of 0.59, representing a well-separated community structure.

The most supported patterns are always representing that the majority of the nodes for each community have the role of non-hub. Although this type of pattern appears in all communities, we can make a distinction in considering the size of the sequence. In Table I, we list the size of some most supported sequential patterns with their community label, community size, sequence size and support. The communities whose sizes are 43 and 41 (i.e. #40 and 77) have long sequences (8 and 7 resp.). Especially, the support of community #77 reaches the maximal value 1. It means there is no remarkable hub author for a long time, or even if they appear sometime, they disappear very quickly in this community. This observation is particularly interesting, and reflects the absence of a community leader who would structure the community through its many connections.

For community #115, the size of the sequence is 1, and its support is also 1.This means all the nodes of this community had the role of non-hub together once, but for the rest of the time slices, they at least took the hub role once. For communities #38 and 40, the support is less than 1, so we can say there is at least one hub, different from the rest of its community, and probably leading it.

TABLE II
Most supported sequence size for each community

Community ID	Community Size	Sequence Size	Support Value
38	335	2	0.99
40	43	8	0.97
77	41	7	1.00
115	125	1	1.00

For community #38, we identify *Philip S. Yu*, *Jiawei Han* and *Beng Chin Ooi* as outliers, i.e. different from their communities. As expected, these nodes have a remarkably high number of connections within their communities, and the represented authors actually have leadership roles in their fields. Further analysis of the data also shows that they publish a total of more than 10 articles per time slice. In addition, they never took the non-hub role.

Let us have a look at the interesting emerging patterns. For community #61, the most emerging pattern is < (ICML PUB. NUM=1) (d is between 3-10 and $z < 2.5$)>, with a growth rate of 3.52 and a support of 0.30. This pattern refers to the authors who published once in ICML, then had a degree between 3 and 10 and became non-hubs. We extract 7 supplementary patterns representing the trends of publishing in AAAI or CIKM to cover all the nodes of this community. The outliers of this community are *Alex Alves Freitas*, *Claire Cardie* , *Edwin P. D. Pednault*. Among these authors *Alex Alves Freitas* does not have any publication for the first 8 time slices, before he starts publishing very frequently in various conferences other than ICML or AAAI and journals. This can be interpreted as a junior researcher progressively maturing. For the other two authors, while *Claire Cardie* publishes in ICML during the first 6 time slices at least once routinely, *Edwin P. D. Pednault* never publishes in not only ICML but also AAAI or CIKM.

The most emerging pattern of community #45 is <(VLDB PUB. NUM=3)(d is between 3-10 $z < 2.5$)> with growth rate 6.40 and support 0.30. This sequence tells us that there is a remarkable group of authors who published 3 times in the VLDB conference, before seeing their degree reach a value between 3 and 10 and holding a non-hub role. The outliers are *Ingmar Weber*, *Anastasia Ailamaki* who do not have any publication for the first 7 time slices, while they both become more and more productive for the last 3 time slices. Their publication number increases fast.

To summarize our observations, the most emerging patterns in almost all communities usually include being non-hub and having a small number of publications in various journals or conferences. Depending on the conferences or journals appearing in these patterns, it is possible to deduce the main theme of these communities. For some communities, however, the emerging sequential patterns are purely topological (no attributes). We can then assume that the members of these communities do not publish in a sufficiently homogeneous way so that it can appear under the form of patterns, which is itself a characteristic of the community. Regarding outliers, one can distinguish different types of profiles. Some seem to correspond to authors whose main theme is different from that of the community in which they were placed. In some cases, we found out the authors had clearly changed their theme, or just started in a given theme. They may also be authors active in another field, including conferences and journals not part of those used in the data we considered here. Another profile is that of junior researchers, whose number of publications and community position evolve jointly. These authors do not seem very active in their field in the first time slices. However, their number of publication and importance in their community increase with time.

V. Conclusions

In this work, we tackled the problem of the characterization of communities in dynamic and attributed complex networks. We proposed a new representation of the information encoded in the network to store the topological information, the node attributes and the temporal dimension simultaneously. We used this representation to perform a search of emerging sequential patterns. Each community could then be characterized by its most distinctive patterns. We also took advantage of the patterns to detect and characterize outlier nodes in each community. We applied our method to a scientific collaboration network constructed from the public database DBLP. The results showed that our method is able to characterize the communities, in particular their research topic. The outlier nodes we identified correspond to different types of profiles, such as community leaders, emerging researchers, or others changing research theme.

Our tool can be improved in various ways. First, its effective processing time could be shortened through various computational optimizations, such as the use of hash maps when calculating growth rates. Second, to limit the complexity of this first approach, we deliberately restrained our analysis method by not considering the evolution of communities over time. In future works, we plan to take advantage of such communities, by inserting the appropriate information in the database used for the search patterns. We also plan to apply our method to the analysis of other types of networks, in order to explore its characterization capabilities. For instance, in our experiment, we saw that there were many nodes whose behavior was not typical of their community. Such observations could be used to study them in further details, and better understand how they are different.

VI. Acknowledgement

This article is supported by the Galatasaray University Research Fund (BAP) within the scope of the project number 12.401.005, and titled “Yabanci Ogrenciler icin Turkce Ogrenim Methodolojisi Gelistirme”. We thank Dr. Sultan N. Turhan, the supervisor of this project.

References

[1] M. E. J. Newman, "The Structure and Function of Complex Networks," *SIAM Review,* vol. 45, pp. 167-256, 2003.

[2] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *PNAS,* vol. 99, pp. 7821-7826, 2002.

[3] S. Fortunato, "Community detection in graphs," *Physics Reports,* vol. 486, pp. 75-174, 2010.

[4] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *JSTAT,* vol. 2008, p. P10008, 2008.

[5] M. Tumminello, S. Miccichè, F. Lillo, J. Varho, J. Piilo, and R. N. Mantegna, "Community characterization of heterogeneous complex systems," *JSTAT,* vol. 2011, p. P01019, 2011.

[6] V. Labatut and J.-M. Balasque, "Detection and Interpretation of Communities in Complex Networks: Practical Methods and Application," in *Computational Social Networks*, ed: Springer London, 2012, pp. 81-113.

[7] Y. Zhou, H. Cheng, and J. Yu, "Graph clustering based on structural/attribute similarities," *Proc. VLDB Endow.,* vol. 2, pp. 718-729, 2009.

[8] J. Yang, J. McAuley, and J. Leskovec, "Community Detection in Networks with Node Attributes," in *ICDM, 2013 IEEE 13th International Conference on Data Mining*, 2013, pp. 1151-1156.

[9] A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato, "Characterizing the Community Structure of Complex Networks," *PLoS ONE,* vol. 5, p. e11976, 2010.

[10] R. Guimerà and L. Nunes Amaral, "Cartography of complex networks: modules and universal roles," *JSTAT,* vol. 2005, p. P02001, 2005.

[11] T. Aynaud and J.-L. Guillaume, "Multi-Step Community Detection and Hierarchical Time Segmentation in Evolving Networks," presented at the SNA-KDD’11, 2011.

[12] X. Yan, J. Han, and R. Afshar, "CloSpan: Mining Closed Sequential Patterns in Large Datasets," in *SIAM SDM '03*, 2003, pp. 166-177.

[13] M. Plantevit and B. Cremilleux, "Condensed Representation of Sequential Patterns According to Frequency-Based Measures," presented at the 8th IDA, 2009.