Tracking Communities over Time in Dynamic Social Network

Etienne Gael Tajeuna[1(✉)], Mohamed Bouguessa[2], and Shengrui Wang[1]

[1] Department of Computer Science, University of Sherbrooke, Sherbrooke, QC, Canada
{etienne.gael.tajeuna,shengrui.wang}@usherbrooke.ca
[2] Department of Computer Science, University of Quebec at Montreal, Montreal, QC, Canada
bouguessa.mohamed@uqam.ca

Abstract. This poster paper presents an approach for tracking community structures. In contrast to the vast majority of existing methods, which are based on time-to-time consecutive evaluation, the proposed approach uses a similarity measure that involves the global temporal aspect of the network under investigation. A notable feature of our approach is that it is able to preserve the generated content across different time points. To demonstrate the suitability of the proposed method, we conducted experiments on real data extracted from the DBLP.

Keywords: Community evolution · Similarity measure · Topological structure

1 Introduction

To understand the evolution of communities over time, several approaches have been proposed. Most of these approaches investigate the common nodes of two communities at consecutive time stamps t_i and t_{i+1} using a Jaccard or modified Jaccard measure [1], [2]. However, as demonstrated in [3], at the end of lifespan such an approach may yield a community that does not share any nodes with the initially observed community.

In fact, a tracking approach that considers only consecutive time points may not necessarily capture the overall temporal evolution of a community. For purposes of clarification, let's look at the the evolution of community $C^1_{t_1}$ from t_1 to t_4 in two different cases as presented in Fig. 1. In the first case (Fig. 1 (First case)) we have an evolution obtained from a simple one-to-one investigation of nodes, with the corresponding evolution of the bag of topics from $B^1_{t_1}$ to $B^1_{t_4}$. As time evolves from time t_1 to t_4, we can see how nodes initially observed in $C^1_{t_1}$ gradually disappear as the topics are gradually change from the computer science field to the mathematic field. However, we can not say that the main topic of community $C^1_{t_1}$ has gradually changed from social network analysis to boolean algebra due to the fact that individuals found in $C^1_{t_4}$ may not share the same

P. Perner (Ed.): MLDM 2016, LNAI 9729, pp. 341–345, 2016.
DOI: 10.1007/978-3-319-41920-6_25

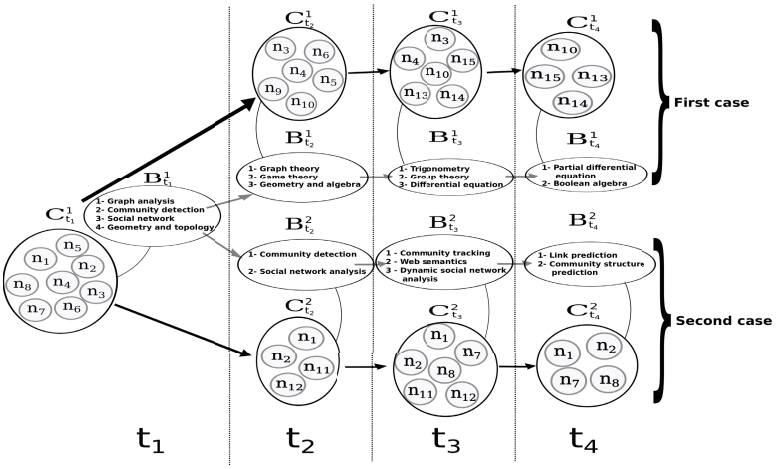

Fig. 1. An example of community and topic evolution.

interest as the individuals initially observed in $C^1_{t_1}$. In the second case (Fig. 1 (Second case)), we have an evolution from $C^1_{t_1}$ to $C^2_{t_4}$, against the corresponding evolution of the bag of topics from $B^1_{t_1}$ to $B^2_{t_4}$. We can see how some nodes initially observed in $C^1_{t_1}$ persist over time, as topics remain in the computer science field due to the fact that individuals found in this evolution are all interested on social network analysis.

The example in Fig. 1 suggests that consecutive tracking yields an inappropriate sequence which may inappropriately reflect the temporal evolution of a community for which the main topics may not be preserved across the time points. To alleviate this, in the next section we present a tracking method that uses a similarity measure which takes into consideration the temporal relation between communities.

2 The Tracking Approach

We denote by $G = \{V_{t_i}, E_{t_i} \,|\, 1 \leq i \leq m\} = (g_{t_i})_{1 \leq i \leq m}$ the dynamic social network over the time period from t_1 to t_m. For each graph g_{t_i}, we define a partition $\left\{C^1_{t_i}, C^2_{t_i}, ..., C^{q_i}_{t_i}\right\}$ representing the communities detected at time t_i using an existing community detection algorithm. To capture the temporal relation aspect of a given community C_{t_j} observed at time t_j, we use the framework that we have developed in [3] to extract the corresponding row vector v_j which captures the temporal aspect relation of C_{t_j} with other communities detected in the social network.

Specifically, we build a binary membership matrix $A_{(N_n \times N_c)}$, in which the rows correspond to the nodes found in G while the columns represent the discovered communities across different time points while. In A, the value 1 indicates that a given node is "present in" a specific community at specific time, while the value 0 reflects the opposite case (that is, the value 0 indicated that a

given node is "absent in" a specific community at specific time). Next, we define the contingency matrix $B = A^T \times A = (b_{\alpha,\beta})_{1 \le \alpha, \beta \le N_c}$, where A^T is the transpose of A . By normalizing each row of B using the relation $p_{\alpha,\beta} = \frac{b_{\alpha,\beta}}{\sum_{\beta=1}^{N_c} b_{\alpha,\beta}}$, we obtain the matrix $B^* = (p_{\alpha,\beta})_{1 \le \alpha, \beta \le N_c}$, where each row j corresponds to the vector v_j. Note that each component $p_{\alpha,\beta}$ represents the probability that community C^α change to community C^β. Hence, the individual row vectors $v_j = (p_{j,1}, p_{j,2}, ..., p_{j,N_c})$ reflect the transition probabilities of community C^j over time which, in turn, represent the proportions of nodes found in community C^j over all detected communities.

To track communities over time, we use the following similarity measure:

$$sim(C_{t_i}, C_{t_j}) = \begin{cases} \sum_{\alpha=1}^{N_c} 2\,\frac{p_{i,\alpha} \times p_{j,\alpha}}{p_{i,\alpha} + p_{j,\alpha}} \;\; if \;\; \sum_{\alpha=1}^{N_c} 2\,\frac{p_{i,\alpha} \times p_{j,\alpha}}{p_{i,\alpha} + p_{j,\alpha}} > \lambda \\ \\ 0 \;\; otherwise \end{cases} \tag{1}$$

where λ is the jonction point between the two Gammas curves estimated from the non-zeros values obtained when scoring the similarity between two transition probabilities vectors; $p_{i,\alpha}$ and $p_{j,\alpha}$ the respective components of vectors v_i and v_j.

By using the similarity measure described in (1), we define the evolution of community C_{t_i}, as the sequence of sorted communities $S_{C_{t_i}} = C_{t_i} \rightarrow C_{t_{i+\eta}} \rightarrow ... \rightarrow C_{t_k}$, $t_i < t_k \le t_m$ such that all communities in $S_{C_{t_i}}$ are similar. Moreover, all communities C_{t_j} in $S_{C_{t_i}}$ should always share nodes with C_{t_i} such that the Jaccard coefficient exceeds zero.

To evaluate the framework presented above, we analyse communities with lifespans of 3 to 14 years. We compare the performance of our approach to existing methods that track communities in a time-sequential manner using similarity measures proposed in [2], [4] and [5]. For an objective comparison of the sequence communities obtained by competing approaches, we adopt a general criterion based on the resemblance between each pair of selected communities in the evolving communities. Specifically, we look at the proportion of nodes persisting in an evolving community (that is, the proportion of nodes observed at the first time and during the time duration of the evolving community). Formally, the proportion of node persisting in an evolving community $S_{C_{t_i}}$, is expressed as follows:

$$N_p(S_{C_{t_i}}) = \frac{1}{|\bigcup_{V \in S_{V_{t_i}}}|} |V_{t_i} \bigcap (S_{V_{t_i}} - V_{t_i})| \tag{2}$$

where $S_{V_{t_i}} = \{V_{t_i}, V_{t_{i+\eta}}, ..., V_{t_j}\}$ is the set of nodes corresponding to the sequence of community $S_{C_{t_i}}$.

Moreover, we look at the general trend of the top 5 most frequent words used for communities with lifespans of 3 to 14 years. Then, we take the particular case of a community with 14 years lifespan to provide a more detailed illustration of the topological structure and content evolution. For topological structure, as time evolves, we look at the transitivity, the conductance [6] and the community's average power to attract and keep nodes [7].

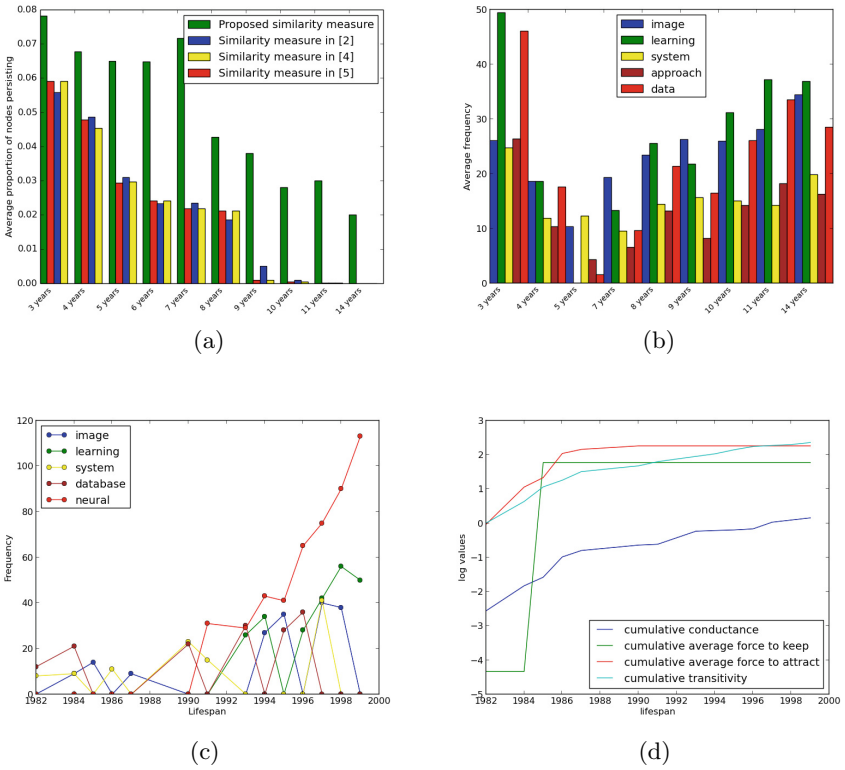

Fig. 2. (*a*) Average proportion of nodes persisting reported by competing approaches. (*b*) General trend of the top 5 most frequent words for communities having lifespans of 3 to 14 years. (*c*) Frequency of the top 5 most used words in $S_{C^{18}_{1982}}$. (*d*) Cumulative topological information.

3 Experiments

We demonstrate the suitability of the proposed approach on the seventh version of the DBLP dataset[1] [8]. The DBLP dataset contains the co-publications of authors. For each paper published it gives the paper title, the authors, the year, the publication venue, the index identification of the paper and the identifications of references to the paper. We built undirected, unweighted graphs between co-authors in the fields of data mining, machine learning and artificial intelligence and databases from year 1977 to 1999, taking each year as a snapshot. Note that in this dataset, nodes correspond to authors and edges reflect co-authorship relations. The total number of nodes is 12,178 while the number of nodes in the different snapshots varies from 80 to 1,709. To identify communities in each snapshot, we use the Infomap algorithm, a parameter-free approach for community detection. For each detected community, we define the bag-of-words obtained from titles of papers written by co-authors.

[1] http://arnetminer.org/citation

The experimental results are given in Fig. 2. Fig. 2(a) gives the histogram of the average proportion of nodes persisting in all competing approaches. As we can see from Fig. 2(a), the average proportion of nodes persisting is higher in our approach. This suggests that the proposed similarity measure is capable to track well communities of co-authors over time than those proposed in [2], [4] and [5]. Moreover, as depicted by Fig. 2(b) which illustrates the general trends of the top 5 most frequent words that persist for communities with lifespans of 3 to 14 years, we can see that, in general, the most frequent words remain present over the different communities' lifespan. The same observation occurs for the community $S_{C^{18}_{1982}}$ with a 14 years lifespan (Fig. 2(c)).

We note that the results observed in Fig. 2(c) can be explained by the cumulative topological structures depicted in Fig. 2(d) where we observe how the capacity of the community to attract and keep a node remains stable from the year 1986, while the conductance and transitivity of their evolution show the strength of the community over time.

4 Conclusion

In this paper, we have presented an approach for tracking community structures in dynamic social networks. Our experiment suggests that the proposed approach can track communities as a sequence and preserve the generated content across different time points. Moreover, the evolution of topological structure explored in the experiment reveals interesting information which may be important to better understand the evolution of communities with their related content. In our continuing research, we are exploring the topological structure to render the approach capable of predicting the future transitions a community may undergo.

Acknowledgments. This work is supported by research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Asur, S., Parthasarathy, S., Ucar, D.: An event-based framework for characterizing the evolutionary behavior of interaction graphs. In: ACM KDD, pp. 913–921 (2007)
2. Takaffoli, M., Fagnan, J., Sangi, F., Zaiane, O.R.: Tracking changes in dynamic information networks. In: IEEE CASoN, pp. 94–101 (2011)
3. Tajeuna, E.G., Bouguessa, M., Wang, S.: Tracking the evolution of community structures in time-evolving social networks. In: IEEE DSAA, pp. 1–10 (2015)
4. Bourqui, R., Gilbert, F., Simonetto, P., Zaidi, F., Sharan, U., Jourdan, F.: Detecting structural changes and command hierarchies in dynamic social networks. In: IEEE ASONAM, pp. 83–88 (2009)
5. Greene, D., Doyle, D., Cunningham, P.: Tracking the evolution of communities in dynamic social networks. In: ASONAM, pp. 176–183 (2010)
6. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Statistical properties of community structure in large social and information networks. In: ACM WWW, pp. 695–704 (2008)
7. Ye, Z., Hu, S., Yu, J.: Adaptive clustering algorithm for community detection in complex networks. Physical Review E **78**, 046115 (2008)
8. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: ACM KDD, pp. 990–998 (2008)