International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 3, March 2016

Tracking Community Strength in Dynamic Networks

[1] K Lakshmi Priya [2] M Pallavi, [3] Prasad B
[1]II/IV, [2][3]Associate Professor
[1][2][3] Department of CSE, Marri Laxman Reddy Institute of Technology and Management (MLRITM) Hyderabad
[1] lakshmipriyakolluru157@gmail.com[2] pallavi_kkreddy@mlritm.ac.in [3]bprasad@gmail.com

Abstract: Analysis on dynamic networks has become a popularly discussed topic today, with more and more emerging data over time. In this paper we investigate the problem of detecting and tracking the variation communities within a given time period. We first define a metric to measure the strength of a community, called the normalized temporal community strength. And then, we propose our analysis framework. The community may evolve over time, either split to multiple communities or merge with others. We address the problem of evolutionary clustering with requirement on temporal smoothness and propose a revised soft clustering method based on non-negative matrix factorization. Then we use a clustering matching method to find the soft correspondence between different community distribution structures. This matching establishes the connection between consecutive snapshots. To estimate the variation rate and meanwhile address the smoothness during continuous evolution, we propose an objective function that combines the conformity of current variation and historical variation trend. In addition, we integrate the weights to the objective function to identify the temporal outliers. An iterative coordinate descent method is proposed to solve the optimization framework. We extensively evaluate our method with a synthetic dataset and several real datasets. The experimental results demonstrate the effectiveness of our method, which is greatly superior to the baselines on detection of the communities with significant variation over time.

***Keywords*: Dynamic, Social network, Community, Tracking, Clustering.**

I. INTRODUCTION

Tracking community strength in dynamic networks has been a hot topic in data mining which has attracted much attention. Recently, there are many studies which focus on discovering communities successively from consecutive snapshots by considering both the current and historical information. However, these methods cannot provide us with much historical or successive information related to the detected communities. Different from previous studies which focus on community detection in dynamic networks, we define a new problem of tracking the progression of the community strength - a novel measure that reflects the community robustness and coherence throughout the entire observation period. To achieve this goal, we propose a novel framework which formulates the Problem as an optimization task. The proposed community strength analysis also provides foundation for a wide variety of related applications such as discovering how the strength of each detected community changes over the entire observation period. To demonstrate that the proposed method provides precise and meaningful evolutionary patterns of communities which are not directly obtainable from traditional methods, we perform extensive experimental studies on one synthetic and five real datasets: social evolution, tweeting interaction, actor relationships, bibliography and biological datasets. Experimental results show that the proposed approach is highly effective in discovering the progression of community strengths and detecting interesting communities. In this, we define that a community is with high strength if it has relatively stronger internal interactions connecting its members than the external interactions with the members to the

rest of the world. Dense internal interactions and weak external interactions guarantee that the community is under a low risk of member change (current members leaving or/and new members joining). Intuitively, a friend community is "strong" if its members tie together closely and ignore the temptation from the outside world. On the contrary, a friend community is regarded as a "weak" community if it is likely to confront a member alteration situation. To illustrate this concept, Fig. 1(a) shows a toy example; where the nodes represented by the same geometric shape belong to the same community, solid lines represent internal interactions and dash lines represent external interactions. The circle community (i.e. nodes A, B, C and D) is considered to be stronger than the rectangle community (i.e. nodes E, F, G and H), due to the weaker external attractions. On the other hand, node H has a close relationship with the diamond community (i.e. nodes I, J and K), which makes the rectangle community in the risk of losing its members.

In other words, the higher strength score a community obtains, the less possible member alternation occurs in it. It is worth noticing that community strength is a measure which synthetically considers both the community cohesion (i.e. how close the members are in a community) and separation (i.e. how distinct a cluster is from the other clusters). Furthermore, community strength should be a temporal measure whose value may change as the network evolves. Here's an example in the real world. A set of authors have collaborated closely from 2000 to 2006. During this period, they cooperated frequently among themselves and barely with others outside the community. However, after 2006, because of interest changes, some authors' attentions have been attracted to some other fields. Thus the internal cooperation decreased and the external cooperation increased. In this case, this author community's strength is high and stable during 2000-2006, but begins to decrease after 2006. As a toy example, in Fig. 1(b) (i.e. the network in the 2ndsnapshot) which evolves from Fig. 1(a) (i.e. the network at the 1snapshot), the strength of the rectangle community decreases, because the internal connections become weaker and external connections become stronger. Discovering the progression of community strengths can offer significant insights in a variety of applications. It can help us discover some interesting community information.

Which cannot be directly obtained from traditional community analysis. Interesting examples of communities' strength progression can be commonly observed in real-life scenarios. Here we discuss two specific cases in detail.

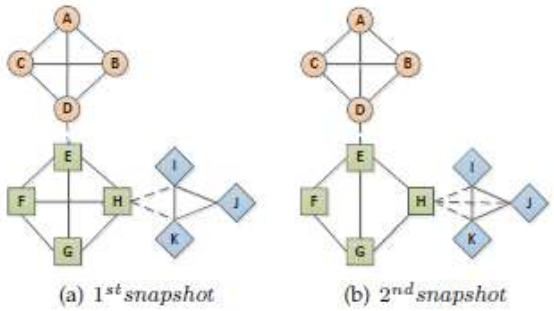

Fig 1: A toy example illustrating community strength

Strengths Progression in Actor Community:
As a strong actor community, the cooperation should be more frequent between the members themselves than between members and non-members. For example, considering the popular and long-running television sitcom 'Friends'1, its six main actors J. Aniston, C. Cox, M. Perry, M. LeBlanc, L. Kudrow and D. Schwimmer collaborated closely when this sitcom was aired from 1994 to 2004. Let's consider each year's co-starring relationships as one snapshot. We can see that the strength of this community is very low before 1994 (little cooperation between them), and then dramatically increases and keeps stable from 1994 to 2004 (average 23 episodes each year). Finally, the strength of this community apparently becomes weaker after 2004 (much less cooperation comparing to the previous years). The progression of this actor community's strengths shows an interesting pattern of cooperation history among these six actors. Learning the strength progression of actor communities helps we better understand the entertainment industry.

Strength Progression in Gene Community:
In the biological domain, the interactions between genes change gradually in dynamic gene co expression networks. Thus the strength of gene communities also changes. For example, it has been reported that the expression profiling of some key genes will change as the cancer progresses. In such cases, the corresponding gene communities' strength also changes. Discovering the strengths of gene communities throughout a specific disease progression can help us find significant clues in the fields of medicine and biology. For a specific disease

if a gene community is found strong only at the early stage, it is very likely to be a crucial trigger for the disease deterioration. From the above cases, we can see that discovering the progression of community strengths helps us understand the underlying behavior of communities. The initial idea was published in which covers the basic definition of community strength and the evolutionary analysis on dynamic networks. By utilizing the community strength value, the consistent communities can be detected and tracked over an observation period.

This paper extends the original idea to formulate a solid method with broader applications and provide more supportive and comprehensive experiments. In this paper, our goal is to detect the temporal strength of each detected community throughout all the snapshots so that we can answer the following questions: How does the strength of each community change over the observation period? What are the top-K strong communities throughout the observation? Period. How do the communities from adjacent snapshots influence the strength of each other? To sum up, our main contributions in this paper are as follows: we introduce the notion of progression analysis of community strengths. To the best of our knowledge, this is the first work on analyzing the temporal community quality or structure information considering both time and community information. We formulate the problem as an optimization framework that can effectively detect the temporal strength of communities and track the strength progression pattern. Experiments on the synthetic dataset show the proposed approach is effective on identifying strong communities. On real datasets, interesting and meaningful communities are detected. Case studies suggest that the proposed approach can provide more reasonable results.

II. METHODOLOGY

The objective of this study is to track the evolution of communities over time in dynamic social networks. We represent a social network by an undirected weighted graph, where the nodes of the graph represent the members of the network, and the edge weights represent the strengths of social ties between members. The edge weights could be obtained by observations of direct inter-action between nodes, such as physical proximity, or inferred by similarities between behavior patterns of nodes. We represent a dynamic social network by a sequence of time snapshots, where the snapshot at time step t is represented by Wt =[wt], the matrix of edge weights at time t. Wt is commonly referred to as the adjacency matrix of the network snapshot. The problem of detecting communities in static networks has been studied by researchers from a wide range of disciplines. Many community detection methods originated from methods of graph partitioning and data clustering. Popular community detection methods include modularity maximization and spectral clustering. In this paper, we address the extension of community detection to dynamic networks, which we call community tracking. We propose to perform community tracking using adaptive evolutionary clustering frameworks, which we now introduce. We present our method for solving the problem of temporal community strength analysis. We begin by introducing the method of partitioning the network from each snapshot into communities in Section 2.1, and then show the method of tracking the strength of each community over time in Section 2.2.

2.1 Community Detection at Each Snapshot

Given a series of temporal networks we first partition each network independently into Kt communities at each timestamp t. Due to the change of network; the value of Kt may not be the same across different snapshots. Then we store all the detected communities from all the snapshots in a community pool. To detect communities from each temporal network, we use Non-negative Matrix Factorization (NMF) techniques there are two major reasons to choose NMF: First, it can be easily applied to both hard clustering (i.e. each object belongs to exactly one community) and soft clustering (i.e. each object can belong to multiple communities). The property of soft clustering very well fits many real social scenarios. For instance, each user in social network usually participates in more than one discussion group, as he may have a variety of interested topics. Second, it could uncover the underlying intercommunity Relationships quite accurately, that can be utilized for other related tasks like progression.

2.2 Tracking communities over time

There are several additional issues that also need to be addressed in order to track communities over time. The communities detected at adjacent time steps need to be matched so that we can observe how any particular community evolves over time. This can be achieved by ending an optimal permutation of the communities at time t to maximize agreement with

those at time t □ 1. If the number of Communities at time t is small, it is possible to exhaustively search through all such permutations. This is, however, impractical for many applications. We employ the following heuristic: match the two communities at time t and t □1 with the largest number of nodes in agreement, remove these communities from consideration, and match the two communities with the second largest number of nodes in agreement, remove them from consideration, and so on until all communities have been exhausted. Another issue is the selection of the number of community's k at each time. Since the evolutionary clustering framework involves simply taking convex combinations of adjacency matrices, any heuristic for choosing the number of communities in ordinary spectral clustering can also be used in evolutionary spectral.

III. EXPERIMENTS

3.1 Reality Mining

Data Description:

The MIT Reality Mining data set was collected as part of an experiment on inferring social networks by monitoring cell phone usage rather than by traditional means such as surveys. The data was collected by recording cell phone activity of 94 students and staff at MIT for over a year. Each phone recorded the Media Access Control (MAC) addresses of nearby Bluetooth devices at minute intervals. Using this device proximity data, we construct a sequence of adjacency matrices where the edge weight between two participants corresponds to the number of intervals where they were in close physical proximity within a time step. We divide the data into time steps of one week, resulting in 46 time steps between August 2004 and June 2005. In this data set, we have partial ground truth to compare against. From the MIT academic calendar, we know the dates of important events such as the beginning and end of school terms. In addition, we know that 26 of the participants were incoming students at the university's business school, while the rest were colleagues working in the same building. Thus we would expect the detected communities to match the participant affiliations, at least during the school terms when students are taking classes.

Observations:

We make several interesting observations about the community structure of this data set and its evolution over time. The importance of temporal smoothing for tracking communities can be seen in Fig. 1. On the left is the heat map of community membership over time when the proposed method is used. On the right is the same heat map when ordinary community detection at each time is used, which is equivalent to setting $\_t$ = 0 in (1). Notice that two clear communities appear in the heat map to the left, where the proposed method is used. The participants above the black line correspond to the colleagues working in the same building, while those below the black line correspond to the incoming business school students. On the heat map to the right, corresponding to ordinary.

3.2 Project Honey Pot

Data Description:

Project Honey Pot is an ongoing project targeted at identifying spammers. It consists of a distributed network of decoy web pages with trap email addresses, which are collected by automated email address harvesters. Both the decoy web pages and the email addresses are monitored, providing us with information about the harvester and email server used for each spam email received at a trap address. A previous study on the Project Honey Pot data found that harvesting is typically done in a centralized manner. Thus harvesters are likely to be associated with spammers, and in this study we assume that the harvesters monitored by Project Honey Pot are indeed representative of spammers. This allows us to associate each spam email with a spammer so that we can track communities of spammers. Unlike in the previous experiment, we cannot observe direct interactions between spammers. The interactions must be inferred through indirect observations. We take the edge weight between two spammers i and j to be the total number of emails sent by i and j through shared email servers, normalized by the product of the number of email addresses collected by i and by j. Since servers act as resources for spammers to distribute emails, the edge weight is a measure of the amount of resources shared between two spammers. We divide the data set into time steps of one month and consider the period from January 2006 to December 2006. The number of trap email addresses monitored by Project. Honey Pot grows over time, so there is a large of new spammers being monitored at each time step. Some spammers also leave the network over time.

IV. RELATED WORKS

There have been several other recent works on the problem of tracking communities in dynamic social networks. Proposed to identify communities by graph coloring; however, their framework assumes that the observed network at each time step is a disjoint union of cliques, whereas we target the more general case where the observed network can be an arbitrary graph. Proposed a method for tracking the evolution of communities that applies to the general case of arbitrary graphs. The method involves rest performing ordinary community detection on time snapshots of the network by maximizing modularity. A graph of communities detected at each time step is then created, and meta-communities of communities are detected in this graph to match communities over time. The main drawback of this approach is that no temporal smoothing is incorporated, so the detected communities are likely to be unstable. Other algorithms for evolutionary clustering have also been proposed. Relevant algorithms for community tracking include and which extend modularity maximization and spectral clustering, respectively, to dynamic data. Proposed an evolutionary spectral clustering algorithm for dynamic multi-mode Networks, which have different classes of nodes and interactions. Such an algorithm is particularly interesting for data where both direct and indirect interactions can be observed. However, one shortcoming in these algorithms is that they require the user to determine to choose the values for parameters that control how smoothly the communities evolve over time. There are generally no guidelines on how these parameters can be chosen in an optimal manner.

V. EXTENSIBILITY TO OTHER APPLICATIONS

By formulating the problem as the task of measuring the community strength over an observation period, we can extend our method to perform some additional tasks. In this section, we explain the extensibility of Algorithm 1to: (1) Measure the impact and consequently the change in community strength based on immediate preceding timestamps, and (2) Identify top-k strongest and weakest communities.

5.1 Community Strength Progression Net

The output of Algorithm 1 provides information on how all the communities' strength evolve over time. In addition to that, we also want to know how the communities from immediate preceding snapshots (i.e. Ct□1 and Ct) influence the strength of each other. To illustrate these relationships, we construct a bipartite network that represents the relationship between communities detected at snapshot t-1 and communities detected at snapshot. In such a network, the nodes on the left represent the communities detected at previous timestamp, the nodes on the right represent the communities detected at the current timestamp and the edges connecting the nodes denote the influence transmission between the communities.

5.2 Top-K strongest/weakest communities

By applying Algorithm 1, we obtain the community strength for each detected community at each snapshot. Based on this output, we can compute an overall strength for each community, which is useful to identify interesting communities that are the strongest/weakest throughout the entire observation period. There are mainly two methods to aggregate the temporal community Strength scores: unweighted and weighted. In the unweighted case, we can regard each temporal score to be of equal importance and take the sum, i.e.PTt=1 azt. However, in some cases, the community strength is more important at some particular snapshots, e.g. the early stage of cancer. In such a case, we should give different weights to different snapshots and the aggregation function Can be defined as PT t=1 htazt, where ht is the weight for the specific snapshot t. In addition, when choosing the top strongest or weakest communities, we may also want to consider the size of the communities. When the target networks are very sparse, the penalty from the external connections may be very small, thus the penalty from the external interaction would be very limited. In such a case, the community strength value would be biased to the large-size communities which will contain more internal connections. To mitigate this effect, the aggregated function for community z can be redefined as:PT t=1 azt jCzj or PT t=1 htazt jCzj so that the community strength is normalized by its size.

VI. CONCLUSION

In this paper, we introduced a method for tracking communities in dynamic social networks by adaptive evolutionary clustering. The method incorporated temporal smoothing to stabilize the variation of communities over time. We applied the method to two real data sets and found good agreement between our results and ground truth, when it was available. We also obtained a statistic that can be used for identifying change points. Finally, we were able to

track communities where the members were continually changing or perhaps assuming multiple identities, which suggests that the proposed method may be a valuable tool for tracking communities in networks of illegal activity. The experiments highlighted several challenges that temporal tracking of community's presents in addition to the challenges present in static community detection. One major challenge is in the validation of communities, both with and without ground truth information. Another major challenge is the selection of the number of communities at each time step. A poor choice for the number of communities may create the appearance of communities merging or splitting when there is no actual change occurring. This remains an open problem even in the case of static networks. The availability of multiple network snapshots may actually simplify this problem since one would expect that the number of communities, much like the community memberships, should evolve smoothly over time. Hence, the development of methods for selecting the number of communities in dynamic networks is an interesting area of future research

REFERENCES

[**1**]. Chi, Y., Song, X., Zhou, D., Hino, K., Tseng, B.L.: Evolutionary spectral clustering by incorporating temporal smoothness. In: Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2007)

[**2**]. Eagle, N., Pent land, A., Laser, and D.: Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences 106(36), 15274{15278 (2009)

[**3**]. Falkowski, T., Bartelheimer, J., Spiliopoulou, and M.: Mining and visualizing the evolution of subgroups in social networks. In: Proc. IEEE/WIC/ACM International Conference on Web Intelligence (2006)

[**4**]. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, and M.W.: Statistical properties of community structure in large social and information networks. In: Proc. 17th International Conference on the World Wide Web (2008)

[**5**].MIT academic calendar 2004-2005, http://web.mit.edu/registrar/www/calendar0405.html

[**6**]. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, and J.P.: Community structure in time-dependent, multistate, and multiplex networks. Science 328(5980), 876{878 (2010)

[**7**]. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577{8582 (2006)

[**8**]. Project Honey Pot, http://www.projecthoneypot.org

[**9**]. Prince, M., Dahl, B., Holloway, L., Keller, A., Langheinrich, and E.: Understanding how spammers steal your e-mail address: An analysis of the rest six months of data from Project Honey Pot. In: Proc. 2nd Conference on Email and AntiSpam (2005)

[**10**]. Tang, L., Liu, H., Zhang, J., Nazeri, and Z.: Community evolution in dynamic multimode networks. In: Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008)

[**11**]. Tantipathananandh, C., Berger-Wolf, T., Kempe, D.: A framework for community identification in dynamic social networks. In: Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2007)

[**12**]. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17(4), 395{416 (2007)

[**13**]. Xu, K.S., Kliger, M., Hero III, A.O.: Evolutionary spectral clustering with adaptive forgetting factor. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (2010)

[**14**]. Yu, S.X., Shi, J.: Multiclass spectral clustering. In: Proc. 9th IEEE International Conference on Computer Vision (2003)