Community evolution in dynamic social networks

Gang Wang
Department of Computer Science and Technology,
Sichuan Police College,
Luzhou Sichuan, 646000, China
E-mail:wgbw2006@sina.com

Yuli Lei, Chongjun Wang
National Key Laboratory for Novel Software Technology,
Nanjing University,
Nanjing Jiangsu, 210093, China
E-mail: chjwang@nju.edu.cn

Shaojie Qiao
School of Information Science and Technology,
Southwest Jiaotong University,
Chengdu Sichuan, 610031, China
E-mail:qiaoshaojie@gmail.com

Abstract-**This paper proposed a framework and an algorithm for identifying communities in dynamic social networks. In order to handle the drawbacks of traditional approaches for social network analysis, we utilize the community similarities and infrequent change of community members combined with community structure optimization to develop a *Group*-based social community identification model to analyze the change of social interaction network with multiple time steps. According to this model ,we introduced a greed-cut algorithm and depth-search-first approach and combine them to develop a new algorithm for dynamic social interaction network recognition (called ADSIN). In addition, we conduct experiments on the dataset of *Southern Women*, the experiment results validate the accuracy and effectiveness of *ADSIN*.**
Key words-social network; community recognition; time step; φ connectitvity graph

I.   INTRODUCTION

Social Interaction Network (SIN) refers to relationship network between individuals in their social activities , some common characteristics in SIN members are often revealed in SIN , such as common blood, common careers, common interests and hobbies, even common illegal crime motive. Investigators in public security authorities tend to lock the parties' social circle or gang members in SIN as soon as possible during the detection process, it has an important practical significance in community identification and its analysis in party-centric SIN.

At present, social community identification method underscoring SIN dynamic is under way of research. The related work includes: group coloring heuristics[1] is a framework to identify a community and its dynamic in its own social network which is based on graph coloring theory with dynamic programming, exhaustive search, maximum match and Greedy Heuristics approach, but this model has some limitation in its real use: its algorithm is too complicated, and it suits for small size social network.

This paper develops a novel algorithm for dynamic social interaction network recognition (ADSIN), as a result of pruning edges and depth search optimization adopted in ADSIN, ADSIN's time complexity is reduced greatly compared with community recognition based on group coloring heuristics model, further more, ADSIN is simple and suits for large social network analysis.

The remainder of this paper is organized as follows. In Section Ⅱ,we introduce some concepts and terminology cited in this paper. Section Ⅲ contains the core of the paper where we present algorithm motivation. The experiment and analysis appears in section Ⅳ. Finally, we present our conclusions in Section Ⅴ.

II.   DEFINITIONS AND PROBLEM STATEMENTS

Definition 1: *Group*
$G=\{x1,x2,…xn\}$ is a set for time step P , entity $xi(i \in [1, n])$ interacts only with entities in set G and doesn't interact with entities in any other sets, then G is *Group*.

Definition 2: *a Set of Group*
G_i (i=1.,..,n)) is a set for time step P, $SG=\{G_1,G_2,…,G_n\}, G_i \cap G_j = \Phi$, where $i \neq j, i,j \in [1, n]$, then *SG is a set of Group or Group set*.

To demonstrate social structure with dynamic setting, assumed that observe a certain amount of social individual in a certain way in a period of time and record the individual interactions at every record time point. Suppose that the input is a group form containing these individuals at every time point. Given an entity set $X=\{x1,x2,…xn\}$, group g is defined as a subset of X,assumed the inputs are a series partition of X: $P_1,P_2,…,P_T$, every partition is a time unit and the intersection between these partition P_i is null. For every time unit, group set defines as $SG_i=\{g_1,g_2,…g_k\}$, where g_1, .., g_k are disjoint entity sets, is defined as an index of partition belong to g. If g belongs to Pi, then $P(g)=i$.

Definition 3: *Time Step for Group Set*
Given a time step $P_1,P_2,…,P_T$ of an entity set X, Sim(,) is a group set similarity measure function, group set $SG=\{SG_1,SG_2,…SG_T\}$, $SG_t=\{g_1,…,g_k\}$ is a group set for every time step t, hence we have:
（1）There is no similarity definitely between any two group for a given time step;   .
（2）In addition to not observeed groups, any one group in subset of group set *SG* for a former time is similar definitely with one group in that of SG for a latter time.
（3）If there is a group *G* in group set *SG* for former time is not similar with any other groups for latter time, then the *G* is a not observed group.

Definition 4: similarity between groups
Given two adjacent time steps *Pi,Pj* and two group sets SG_i,SG_j corresponding to time steps *Pi,Pj* respectively, and $g_i \in SG_i, g_j \in SG_j$, if g_i is similar to g_j, then

Sim(g$_i$,g$_j$)=max(Sim(g$_i$,SG$_j$)), where *max(Sim(g$_i$,SG$_j$))* is the maximum similarity between gi and all groups in SGj .

III. ALGORITHM IDEA & DESCRIPTION

A. Algorithm Idea

The proposed model is expressed by a right-weighted and multi-division directed graph as G=(V, ω ,E), the V express the nodes of the graph, the ω expresses the right weigh and E expresses the relationships among the nodes. The Specific rules are as follows:

(1) The formula V={V$_1$,…V$_T$}，V$_i$（i∈［1, T］） is the group set SGi in the time step Pi. If the time steps P(gi) and P(gj) are adjacent and Sim(gi,gj)=max(Sim(gi,SG$_j$)), then (gi,gj)∈E.

(2) The right weight ω (gi,gj)=Sim(gi,gj).

(3) E={E1,…,Ek}，Ei={ei(Vx,Vy)|Vx∈V∧Vy∈V ∧Vx≠Vy}（i∈[1,K]），the E$_i$ expresses the relationship i of the graph.

B. Hypothesis and Measurements

In order to measure the group members' aggregations, we assuming the individuals aggregation for the first time to this group are 0 and we propose the three measurement elements: in-degree value, out-degree value and constant value. α,β,γ are used to calculate the individuals interactions in the group.

In-degree value: Supposing there is an edge from gt to gt+1 , if an individual x at time step t does not belong to group gt, but belongs to group gt+1 at time step t+1, then we get the aggregation of individual values x decreased by α .

Out-degree value: Supposing there is an edge from gt to gt+1 , if an individual x belongs to group gt at time step t, but not belongs to group gt+1 at time step t+1, then we get the aggregation of individual values x decreased by β.

Constant value: Supposing there is an edge from gt to gt+1 , if an individual x belongs to group gt at time step t, and belongs to group gt+1 at time step t+1, then we get the aggregation of individual values x increased by γ.

C. Algorithm Description

Based on the above-mentioned definitions, ADSIN algorithm is as follows:

Step1 Determine the time observing point, that is, dividing the time steps, and statistic the group set for each time step.

Step2 Calculate the similarity of groups between the adjacent time steps, and build relationships set model graph algorithms.

Step3 Do the pruning treatment to the model graph, when some nodes for the previous time step point to the same node the next time step, select the largest similarity associated node and retain it, cut the other nodes associated with the node.

Step4 Find the maximum connected graph which path

length is greater than φ , use depth-search-first approach at this step.

Step5 Calculate each group member aggregation of the connected graph, if a group member's aggregation> ξ, then it is a real member of the group, finally we can obtain a group clustering. Removed Just connected graph from the graph, we jump back to **Step4**, until there is no connected graph which maximum path length is greater than φ, the algorithm ends.

D. Time Complexity Analysis

We make T as the total number of time steps, n as the number of group sets for each time step, m as the member number for each group set. According to Algorithm Description above mentioned, we can analyze that the computational complexity of the algorithm which calculated the similarity of groups between adjacent time steps is O (Tmn^2); Greed-cut algorithm will traversal all the vertices of the group, the complexity is O(nT); The depth-search-first approach can find the maximum connected graph which the path length is larger than φ, for the depth-search connect graph is beginning to search the group vertices which don't have in-degree value, so its complexity is O(nT).

IV. EXPERIMENTAL RESULTS

A. Experiment Results

We apply the ADSIN algorithm to the southern women data set[2], the result is shown in Figure 1.
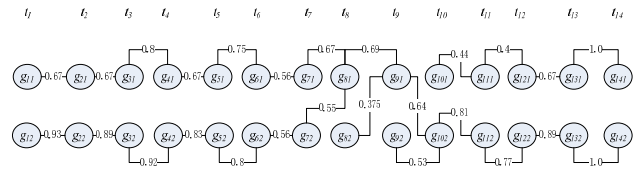

Figure 1. Dynamic Social Network Model Used by ADSIN's Algorithm

From Figure 1, there exist two group vertices for the previous time step pointing to the same vertices for the next time step in the model graph, and that is to say, we should optimize the model graph. We can obtain the model graph shown in Figure 2 with the cutting algorithm. In the optimized graph, besides one single node, there leave three community groups.
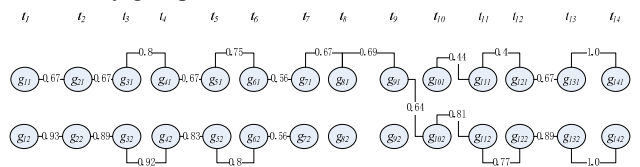

Figure 2. Dynamic Social Network Model After Optimized Community Structure

B. Experiment Results Analysis

At first, we set the parameters as $(\alpha, \beta, \gamma) = (1, 1, 3)$, the aggregation ξ =1, connected graph coefficient φ =1, then we can figure out the aggregations for all the

community group members, which is shown in Table Ⅰ, and this result will also display all the members and communty groups.

Secondly, we set the aggregation ξ and connected graph coefficient φ again. To exclude the fake members and get all the real community members, we need set ξ =12 and φ =6. That setting can exclude some fake members and community groups and the result is shown in Table II, which only display the members and the community groups whose parameters are larger than given parameters.

Finally ,we can see that the aggregation of No.16 between two communities is nearly equal in the Table II . So we can set definitely the entity belongs to the two communities during the experiment time. This is the overlapped community structure, which do not against the real social structure. Moreover, we also can see some group members belong to neither of the two communities like the member number 6. Its aggregation is 8 and it only appear in community 1. But it is excluded out for its small aggregation, which means it wandering around the community 1, so we could not make sure which community it belongs to.Seeing from Table II,we can get community recognition results in Table III.

Table Ⅲ. Result Sets for Community Member Recognition

Group Number	Group Members
1	1,2,3,4,5,7,8,9,16
2	10,11,12,13,15,16,17,18

As the Table III shows, we can obtain that this result of the proposed algorithm is similar with the result given by Freeman[3], however, there still exist differences that entity number 6 is belonging to the community 1 with 21kinds of different social networks recognition algorithms, but based on the proposed algorithm, the member is wandering around community 1 and need to be observing. That is, this algorithm result is definitely fitting the fact.

V.   CONCLUSIONS

This paper analyzed the disadvantages of traditional social network recognition algorithm. In order to characterize the dynamic of reality social activities, we defined groups, group sets, time steps for group sets, and the similarity for the groups set; with community structure optimization algorithm, we used Group-based social network recognition model to analyze the change of social network recognition in the multi-time steps and then based on this model, we proposed a new social dynamic networking algorithms, ADSIN of the social network recognition, the innovation of the algorithm are as follows:

(1) a Novel group-based social network recognition model.

(2) With the greed-cut algorithm and depth-search-first approach, we optimized the community structure and reduce the computational complexity.

Compared this model with the former dynamic model, we reduce the complexity with greed-cut algorithm and depth-search-first approach, and this model can be applied in lagre social network analysis. Based on the verification on the *Southern Women* classical data sets, This algorithm has more accuracy and validity,and the analysis of the results shows more objectively.

In terms of our future work, we will employ some new data mining approaches that has been proposed in literatures [6-9] to improve the predict accuracy of the proposed approach.

References

[1] C. Tantipathananandh,T. B. Wolf and D. Kempe, A framework for community identification in dynamic social networks，Knowledge Discovery and Data Mining，KDD'07.

[2] A. Davis, B. B. Gardner, and M. R. Gardner. Deep South. The U. of Chicago Press, Chicago, IL, 1941.

[3] L. Freeman, Finding social groups: A meta-analysis of the southern women data, Dynamic Social Network Modeling and Analysis--workshop summary and papers, The National Research Council(National Academies Press), Washington, D.C., 2003.

[4] Lisa Singh,Mitchell Beard, Lise Getoor.Visual Mining of Multi-Modal Social Networks at Different Abstraction Levels. 11th International Conference Information Visualization ,2007.

[5] Matteo Brunelli, Michele Fedrizzi.A fuzzy approach to social network analysis. Advances in Social Network Analysis and Mining. 2009.

[6] Shaojie Qiao, Tianrui Li, Jing Peng, Jiangtao Qiu. Parallel Sequential Pattern Mining of Massive Trajectory Data. International Journal of Computational Intelligence Systems, 3(3): 343-356, 2010.

[7] Shaojie Qiao, Changjie Tang, Huidong Jin, Teng Long, Shucheng Dai, Yungchang Ku, Michael Chau. PutMode: Prediction of Uncertain Trajectories in Moving Objects Databases. Applied Intelligence, 33(3): 370-386, 2010.

[8] Shaojie Qiao, Changjie Tang, Huidong Jin, Shucheng Dai, Xingshu Chen. Constrained K-Closest Pairs Query Processing Based on Growing Window in Crime Databases. In: Proceedings of 2008 IEEE International Conference on Intelligence and Security Informatics, ISI 2008, Taipei, pp: 58-63, 2008.

[9] Shaojie Qiao, Changjie Tang, Huidong Jin, Jing Peng, Darren Davis, and Nan Han. KISTCM: knowledge discovery system for traditional Chinese medicine. Applied Intelligence, 32(3): 346-363, 2010.

Proceedings of the 2012 2nd International Conference on Computer and Information Application (ICCIA 2012)

Table I. Results 1 for Community Recognition

Group Number	Group Members	aggregation	
		Member Number--Value	Member Number--Value
1	g11,g21,g31,g41,g51,g61,g71,g81,g91,g102,g112,g122,g132,g142	key-value:3--35 key-value:1--31 key-value:2--23 key-value:4--23 key-value:7--18 key-value:9--18 key-value:16--17 key-value:8--15 key-value:10--14 key-value:7--8	key-value:5--13 key-value:10--12 key-value:17--6 key-value:18--6 key-value:15--3 key-value:13--2 key-value:11--1 key-value:6--8 key-value:8--10 key-value:16--18
2	g12,g22,g32,g42,g52,g62,g71	key-value:17--18 key-value:18--18 key-value:9--6 key-value:15--14	key-value:13--14 key-value:14--11 key-value:11--18 key-value:12--18
3	g101,g111,g121,g131,g141	key-value:15--5 key-value:13--4	key-value:14--12 key-value:12--4

Table II. Results 2 for Community Recognition

Group Number	Group Members	aggregation (Member Number--Value) ξ =15	aggregation (Member Number--Value) ξ =13
1	g11,g21,g31,g41,g51,g61,g71,g81,g91,g102,g112,g122,g132,g142	key-value:1--31;key-value:2--23 key-value:3--35;key-value:4--23 key-value:7--18;key-value:8--15 key-value:9--18;key-value:16--17	key-value:3--35;key-value:2--23 key-value:1--31;key-value:7--18 key-value:5--13;key-value:4--23 key-value:9--18;key-value:8--15 key-value:16--17;
2	g12,g22,g32,g42,g52,g62,g71	key-value:17--18;key-value:18--18 key-value:16--18;key-value:11--18 key-value:12--18	key-value:10--14;key-value:17--18 key-value:18--18;key-value:15--14 key-value:16--18;key-value:13--14 key-value:11--18;key-value:12--18