PROGRESSIVE K-CLIQUE MINING IN CO-AUTHORSHIP NETWORK BASED ON TEMPORAL CONCEPT ANALYSIS

V.Akila
Pondicherry Engineering College, Pondicherry,India
akila@pec.edu

V.Govindasamy
Pondicherry Engineering College, Pondicherry,India
vgopu@pec.edu

Abstract - Social Network is omnipresent in today's world .Co-authorship network is a subset of Social Network. Interactions in Co-authorship network are based on common interests or similar profiles. The formation of groups is unavoidable in this setting. The identification of these groups in this setting may help in harnessing the social dynamics that exist in the institution. This will in turn assist in understanding the informal research groups evolution in an institution. It should also taken to consideration that Co-authorship networks usually evolve over time. It is a complex task to efficiently identify k-cliques from dynamic social networks. To address this challenge, this paper proposes an efficient k-clique detection method based on Temporal Concept Analysis (TCA). Experimental results illustrate that the proposed detection method is efficient for extracting the k-cliques from the Co-authorship networks.

Keywords**:** Social Network; Temporal Concept Analysis; k-cliques

1. Introduction

Co-authorship Network is a special kind of Social Network. They model the collaboration of authors in research publications. This network essentially encode latent details about the collaboration behavior of authors in an organization. Understanding this implicit behavior of authors can assist in the understanding of the informal subgroups that transcend departments, subject area etc. Co-authorship Networks like Social networks contain collections of sub-networks . If the largest component contains a majority of nodes in a network, it is known as a maximal sub graph. Analysis of Co-authorship Networks leads to the identification of subgroups of authors who are more closely linked to one another in "communities" or "cliques" . The k-clique detection problem is a basic problem in computer science that can help understand the behavior patterns of authors in Co-authorship networks [Hao et al.(2017), Hao et al.(2016)].

A Co-authorship network, where the graph's vertices represent the authors, and the graph's edges represent the authorship of papers. Then a clique represents a subset of people who have published together, and algorithms for finding cliques can be used to discover these groups of authors. Clique detection is one of the most common problems in social network. At present, there is no efficient way to find k-clique in a dynamic social network .To address this problem, we find an efficient method to detect k-cliques which can be used to understand the behavior among authors and to identify an exclusive circle of people with a common research interest.

Inspired by the properties of Temporal Concept Analysis and dynamical features of social network, this paper aims at discovering k-clique in dynamic social network efficiently by extending the Formal Concept Analysis (FCA) by adding time dimension to it. The detected k-clique helps in discovering some useful structures to understand the users behaviors in dynamic social network. This paper analyses the extracted k-cliques progressively using the CliqueAnalyser algorithm.

2. Literature Review

The Literature review concentrates on the existing k-clique mining techniques.

Clique Percolation Method (CPM) defines a model for extracting a k-clique template. The CPM is usually applied for analyzing the overlapping community structure in networks. Clique Percolation Method (CPM) [Palla et al.(2008)] defines a k-clique-community as an amalgamation of adjacent k-cliques. The limitation of the paper is that i) CPM time estimates are not based on statistical analysis. ii) CPM fails to propose the object function for quantitative qualification of the clustering result. iii) CPM fails to capture the community structure of sparse networks.

The Sequential Clique Percolation(SCP) [Kumpula et al.(2008)]algorithm is useful for the following three cases i) when 'k' is used ii) for multiple weight threshold levels iii) no apriori knowledge of the threshold level of dense weighted network. The algorithm performs well for very large sparse networks. The limitations of the paper are i) Poor performance on networks with pervasively overlapping community structure ii) It cannot generate k-clique communities for various 'k' values in a single execution.

The paper [Pollner(2012)] makes use of CFinder and CPM to locate the k-clique percolation clusters of the network. It is used in bioinformatics. It has been proved that CFinder can be used to predict the function of a single protein as well as to discover novel modules. CFinder is efficient for locating the cliques of large sparse graphs. The limitations of the paper are that it cannot find clique in densely interconnected nodes in graphs.

Com Tector[Baum (2003)] first enumerates all maximal cliques in the giant component of a social network. It is reasoned that a maximal clique is a complete sub-graph. Because of this it is the densest community. This can be used to represent the closest relationship involving a single entity in the given network. This can be leveraged using the overlapping nature of the communities in real world scenarios.

The eagle algorithm [Shen et al.(2009)] is used . This algorithm deals with maximal cliques within an agglomerative framework. Communities from Edge Structure and Node Attributes (CENSA) [Shen et al.(2009)] is proposed for detection of detecting overlapping communities in networks. It has been proved to be an accurate and scalable. It provides a statistical model for interaction between network structure and node attributes. This modeling purports to more accurate community detection. It provides improved robustness in the presence of noise in the network structure. CESNA also helps with the interpretation of detected communities by finding relevant node attributes for each community.

The limitations inferred from the survey is though there are a number of methods existing to extract k-cliques from static graphs there is a paucity of methods to extract stable k-cliques from evolving graphs. To this end, this paper presents a method that is based on Temporal Concept Analysis and proposes a new algorithm CliqueAnalyser which progressively extracts the k-cliques.

3. K-Cliques with Temporal Concept Analysis

3.1. *Problem Definition*

Definition 1 (k-clique).

Let G = (V, E) be an undirected graph. A k-clique in G is a subset S ⊂V and |S| = k such that for any two vertices v1, v2 ∈ S there exists an edge (v1, v2) ∈ E

The problem is to find k-clique in dynamically varying social network.

3.2. *CliqueAnalyser*

Temporal Concept Analysis is an extension of Formal Concept Analysis (FCA)[Poelmans et al.(2010), Wolf (2011) , Zhang et al.(2014)]. It introduces a time component to concept lattices allowing concepts to evolve with time . The evolution of concepts in a dynamic network can assist in understanding the dynamics of the network. Social Network Analysis has been done through Formal Concept Analysis. Extending this theory, Co-Authorship Network can be converted to Temporal concept T=(V,V,I,t) where 'V' is the authors in the graph and 'I' is the relation that defines if an author has collaborated with the other author at time 't' [Hao et al.(2017), Hao et al.(2016)] . The time 't' is based on the division of each year as 1st quarter,2nd quarter, 3rd quarter and 4th quarter of an year. For the given network, the temporal concepts are generated. For the initial quarter of the year of publications the concept lattice 'F' is built. The extent-extent overlap matrix is constructed. The k-cliques are found equivalent to extent components in the k-intent concepts. The clique matrix 'M' consisting of the k-cliques is saved .The temporal concepts from the next quarter and the aforementioned steps are performed iteratively. The resultant k-clique matrixes are saved. The algorithm for extracting the most stable cliques over time is given in Figure.1. The difference between the successive clique matrixes are examined. The sign of each element in the difference matrix is investigated. If the sign is negative then it indicates that an previously existing relation is missing. Then the weight between the authors are reduced by a depreciating factor 'df'. The depreciation is done in a manner to model gradual forgetting [Akila et al.(2016)]. This is to factor in that groups of recent times should have more weight than the groups in the past. The earlier the relation is dissolved, the deduction of weight is more than if the relation is dissolved in recent times. Similarly any relation in the clique that reappears repeatedly is awarded with a weight 'wt'.

```
Algorithm CliqueAnalyser
Input:
  Set of matrixes from each iteration of the K-clique construction
Output:
  Set of k-cliques
Begin
  Initialize k-set as NULL
  For i=1 to i= last-1 iteration do
    Begin
      Temp=Mi+1-Mi
      Mi+1=Mi+1+Mi
      For each element in the matrix where j is the row and  k is the col do
      Begin
        If Temp[j,k]<0
          Mi+1[j,k]= Mi[j,k]*(1- (df/i))
        Else
          Mi+1[j,k]= Mi+1[j,k]*(1+(i/wt)) + Mi[j,k]
      End
    End
End
```

Figure 1: Algorithm CliqueAnalyser

4. Experimental Evaluation

The environment used for the development is a Pentium IV 2.4 GHz with a HDD 40 GB and RAM 512 MB. The system used a Windows XP Operating System and the platform used was .NET. The Front End is ASP.NET 3.5, Java and Backend is SQL SERVER 2008.The Data set used is a Co-authorship network of publications in an academic institution from the year 2011 to 2015. Each year was segmented to four quarters.The number of iterations are 20. The results of the experimental analysis is given in Figure 2.The value for ‘k’ was set as 4 since according the rule of thumb the average Co-authors of a paper is 4. From the graph in Figure.2., it is clear that the existing methods based on Formal Concept Analysis the number of cliques are increased as the size of the graph is increased with time. But, it does not delineates between active and in active groups. Whereas the proposed system extracts the stable cliques in an evolving scenario.

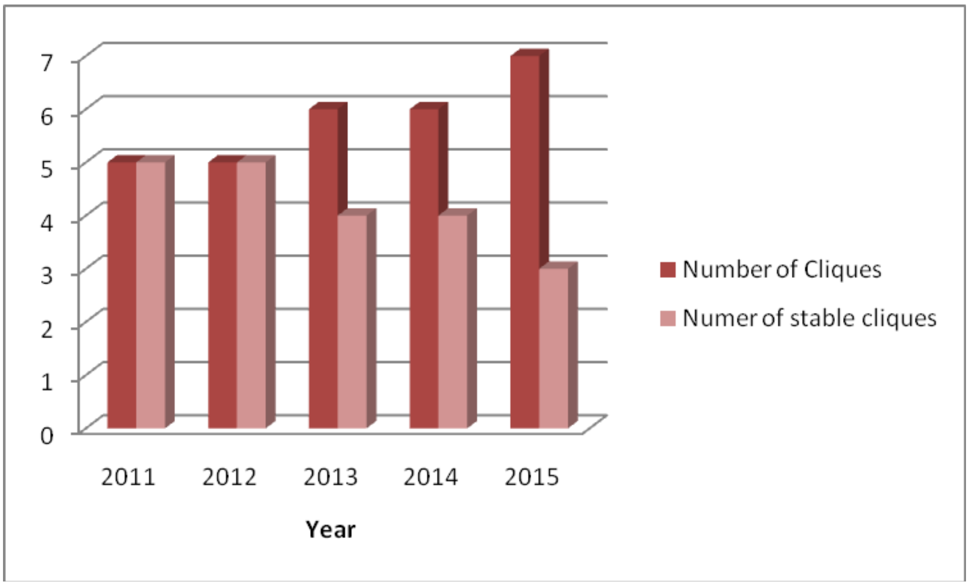

Figure 2: Performance evaluation of the CliqueAnalyser algorithm

5. Conclusion

Clique is a very common structure in social networks. It is composed of the set of vertices as well as the mutual relationships among them. This reflects the social behavior and its social features among users. This kind of study is essential in an academic setting to understand the dynamics of collaboration among the researchers. This paper proposes a method based on Temporal Concept Analysis and CliqueAnalyser algorithm to identify stable k-cliques in a Co-authorship network. The experimental analysis reveal that method based on simple Formal Concept Analysis extracts the k-cliques, but these methods are not efficient in a dynamic setting. Future work can involve finding frequent k-clique in the dynamic network.

Acknowledgments

The authors acknowledge the support of S. Leela, L.Nishanthini and A.ShanthaMeena of Department of Computer Science and Engineering, Pondicherry Engineering College in their assistance of this work. This work was supported by the UGC Minor Research Project with Proposal Number:2477(F.NO:4-4/2015-16(MRP/UGC-SERO).

References

[1] Baum D. Finding all maximal cliques of a family of induced subgraphs. Konrad-Zuse-ZentrumfürInformationstechnik Berlin [ZIB]; 2003.

[2] Hao, F., Min, G., Pei, Z., Park, D.S. and Yang, L.T., 2017. $ K $-Clique Community Detection in Social Networks Based on Formal Concept Analysis. IEEE Systems Journal, 11(1), pp.250-259.

[3] Hao, F., Park, D.S., Min, G., Jeong, Y.S. and Park, J.H., 2016. k-Cliques mining in dynamic social networks based on triadic formal concept analysis. Neurocomputing, 209, pp.57-66.

[4] Kumpula, Jussi M., MikkoKivelä, KimmoKaski, and JariSaramäki. "Sequential algorithm for fast clique percolation." Physical Review E 78, no. 2 (2008): 026109.

[5] Palla, Gergely, DánielÁbel, Illés J. Farkas, PéterPollner, ImreDerényi, and TamásVicsek. "k-clique Percolation and Clustering." In Handbook of Large-Scale Random Networks, pp. 369-408. Springer, Berlin, Heidelberg, 2008.

[6] Poelmans, J., Elzinga, P., Viaene, S., Dedene, G.: A Method based on Temporal Concept Analysis for Detecting and Profiling Human Trafficking Suspects. In: Artificial Intelligence and Applications, AIA 2010, Innsbruck, Austria. pp. 1–9 (2010)

[7] Pollner P, Palla G, Vicsek T. Parallel clustering with CFinder. Parallel Processing Letters. 2012 Mar;22(01):1240001.

[8] Shen H, Cheng X, Cai K, Hu MB. Detect overlapping and hierarchical community structure in networks. Physica A: Statistical Mechanics and its Applications. 2009 Apr 15;388(8):1706-12.

[9] V. Akila, V. Govindasamy, S.Sharmila, " Bug Triage Based on Ant System with Evaporation Factor Tuning", International Journal of Control Theory and applications, 9(2) 2016, pp. 859-863

[10] Wolff KE. Temporal concept analysis explained by examples. CDUD'11–Concept Discovery in Unstructured Data. 2011:104.

[11] Yang J, McAuley J, Leskovec J. Community detection in networks with node attributes. InData Mining (ICDM), 2013 IEEE 13th international conference on 2013 Dec 7 (pp. 1151-1156). IEEE.

[12] Zhang Y, Wang X, Liu J. Event dynamic transition model in TCA based on the attribute partial diagram. InControl and Decision Conference (2014 CCDC), The 26th Chinese 2014 May 31 (pp. 5012-5016). IEEE.