International Journal of Information Technology and Knowledge Management
January-June 2011, Volume 4, No. 1, pp. 43-46

# STATISTICAL APPROACH OF SOCIAL NETWORK IN COMMUNITY MINING

Meenu Gupta[1] & Rajeev Yadav[2]

The popularity of social networking on the web and the explosive combination with data mining techniques open up vast and so far unexplored opportunities for social intelligence on the Web. A network community is a special sub-network that contains a group of nodes sharing similar linked patterns. A social network can be defined as a graph G = (V, E), where V = {$v_1$, $v_2$, $v_3$, ….,$v_n$} is the set of vertices, and E is the set of edges connecting pairs of vertices. Each edge represents the social relationships among two nodes representing people. Analyzing the social network can help gain further understanding on the characteristics of the social networks. Many community mining algorithms have been developed in the past. In this work, we have presented a new algorithm BFC which uses statistical approach for community mining in Social networks. The algorithm proceeds in breadth first way and incrementally extract communities from the Network. This algorithm is simple, fast and can be scaled easily for large Social networks. The effectiveness of this approach has been validated using implementation in GUESS (Graph Exploration System) tool and network examples.

## 1. Introduction

Many scientific and commercial applications needs patterns that are more complicated then frequent item-set and sequential patterns and require extra effort to discover. Such sophisticated patterns go beyond sets and sequence, towards trees, lattices, graphs, networks, and other complex structures. Such complex networks are extensively studied on Social Networks.

Social Networks include online community networks, disease transmission networks, corporate executive networks, criminal/terrorist networks etc. Social Network Analysis (SNA) is one of the emergent fields of research for extracting useful information from social network data. Real-life communities are formed by people working together, sharing a hobby, living nearby each other, etc. Community structure is an important topological property of social networks which could provide a higher logical view of network, and will dramatically decrease the dimensionality while analyzing the structure and evolution of the social network. Many community mining algorithms have been developed in the past but there exists many problems making them perform slower and inefficient.

### 1.1. Motivation

Social Networks have evolved much and became hot topic due to lot of applications and research in this field. In the literature, many algorithms have been developed to detect network communities or sub-graph clustering. They can generally be divided as in [1] into three main categories:

1. Graph theoretic methods like Random walk methods and physics-based methods Spectral methods
2. Divisive algorithms like ‘Betweenness’ algorithms of Girvan and Newman [24] Tyler algorithm[23] and Radicchi algorithm [21] in which they divide the network into smaller subsections
3. Agglomerative algorithms like Modularity-based algorithms [18] which form communities by joining nodes together.

Girvan and Newman [24] proposed ‘betweenness’ measure in 2002 which iteratively removes edges with the highest “stress” to eventually find disjoint communities. Clauset [18] in 2004 suggested a faster algorithm but the number of clusters must still be specified by the user. Flake et al. [25] in 2000 used max-flow min-cut formulation to find communities around a seed node; however, the selection of seed nodes is not fully automatic.

Kelsic [12] in 2005 proposed an agglomerative algorithm for constructing overlapping communities using local shells, and implement methods for visualizing overlap between communities. Pons and Latapy [13] in 2005 reported a community finding method using random walk. It starts with single-node communities and repeatedly performs the merging of a pair of adjacent communities that minimizes the mean of the squared distances between each node and its community.

[1]Student of M.Tech (4th sem),

[2]Research Scholar Dravidian Univeristy Kuppam

Email: [1]gupta.meenu5@gmail.com, [2]rajeevtpo@gmail.com

Hildrum [10] in 2005 presented a cut-based focused community search algorithm. Palla [14] in 2005 used clique percolation for the problem of identifying communities, where one node can belong to more than one community. They viewed a community as a union of all k-cliques (that is, complete subgraphs of size k) and studied the statistical features (for example, the cumulative distributions of the community size, community degree or number of overlap links, overlap size, and membership number or number of communities) of the interwoven sets of overlapping communities involving highly overlapping cohesive groups of nodes. Their method first identifies all cliques of the network and performs a standard component analysis of the clique-clique overlap matrix to discover a set of k-clique-communities.

Kim and Jeong [15] in 2005 developed a matrix block diagonalization and applied it to weighted stock networks. Their method constructs a network of stocks and identifies stock groups with a percolation approach based on a filtered empirical stock correlation matrix.

Newman [9] in 2006 introduced eigen-spectrum of a matrix and calls it modularity matrix, which plays a role in community detection similar to that played by the graph Laplacian in graph partitioning calculations. Qian [6] in 2006 presented a link mining algorithm to identify communities of practice based on the idea that linked nodes belonging to the same community should have a larger number of 'common friends'. Ichise[7] in 2006 presented, a community mining system which helps to find communities of researchers by using bibliography data, in this method the key feature is the modeling of papers and researchers, which enables us to eliminate the edges of large clusters.

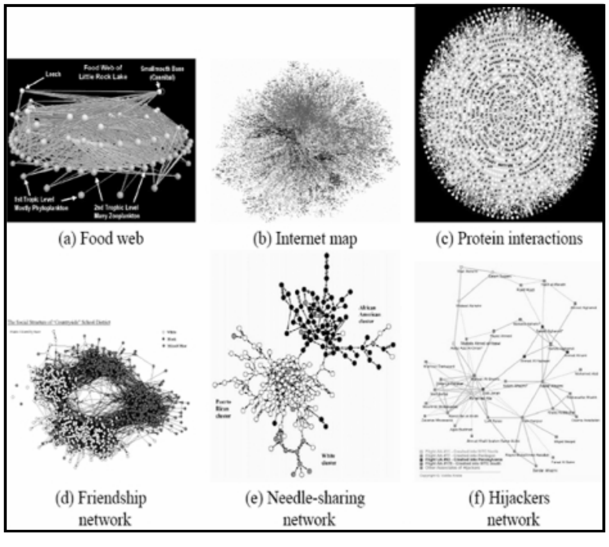
Yang and Liu [5] in 2006 presented an incremental force-based algorithm which allows mining communities in large scale dynamic networks, which is inspired by Newtonian gravitational law; it considers degree of vertex as mass and each edge as virtual spring. Recently Yang et al. [1] in 2007 developed a new algorithm, called FEC, for identifying communities from signed social networks. The key idea behind it rests on an agent-based random walk model, based on which the FC phase can find the sink community including a specified node with a linear time complexity. Thereafter, the sink community is extracted from the entire network by the EC phase based on some robust graph cut criteria.

In one other paper by Yang and Liu [2] in 2007 they presented an Agent-based AOC approach to solving Distributed Network Community Mining Problem (D-NCMP), In this approach, the nodes and links of distributed networks are distributed among a group of autonomous agents, who are responsible for finding all natural communities hidden in distributed networks, based on their respective local views.

## 2.1. Social Networks

A Social Network comprises of social structure of nodes tied together with one or more type of relationship such as friendship, dislike, trade, financial exchange, etc. A social network can be defined as a graph G = (V, E), where V = {$V_1$, $V_2$, $V_3$, ….,$V_n$} is the set of vertices, and E is the set of edges connecting pairs of vertices. Each edge represents the social relationships among two nodes representing people. They have heterogeneous and multi-relational dataset represented by graphs. Typically these graphs are very large and both nodes and links have attributes. Social networks need not to social in context. There are many real world instances of economic, biological, technological and business social networks.

Social Networks include online community networks, electrical power grids, disease transmission networks, corporate executive networks, the spread of computer viruses, criminal/ terrorist networks etc. Customer networks and collaborative filtering problems (where product recommendation is made based on the preferences of other customers) are other examples. In biology, examples range from epidemiological networks, cellular and metabolic networks, and food webs, to the neural network of the nematado worm Caenorhabditis elegans(the only creature whose neural network has been completely mapped). The exchange of email messages within corporations, newsgroups, chat rooms, friendships, and the quintessential "old boy" network (i.e., the overlapping board of directors of the largest companies in the United States) are examples from sociology.


Leech
Food Web of Little Rock Lake
Smallmouth Bass (Cannibal)
1st Trophic Level Mostly Phytoplankton
2nd Trophic Level Many Zooplankton

(a) Food web (b) Internet map (c) Protein interactions

(d) Friendship network (e) Needle-sharing network (f) Hijackers network

Fig. 2.1: Example of Graphs that can be Modeled into Social Networks

### 2.2. Comparison with Existing Algorithms in Literature

Table 2.1
Comparison with Algorithms in Literature

| S. No. | Algorithm | Type | Order of Time Complexity | Fully automatic |
|---|---|---|---|---|
| 1 | Newman Betweenness | Divisive | $O(N^3)$ | yes |
| 2 | Max Flow Min Cut | Divisive | $O(N \log_2 N)$ | yes |
| 3 | Modularity | Agglomerative | $O(N(M + N))$ | yes |
| 4 | Improved Modularity | Agglomerative | $O(N \log_2 N)$ | yes |
| 5 | Spectral Partitioning | Divisive | $O(N^2)$ | no |
| 6 | External Optimization | Divisive | $O(N^2 \ln N)$. | no |
| 7 | Force Based theoretic | Graph | $O(MxN^2)$ | yes |
| 8 | Link mining Based | Divisive | $O(N^2)$ | yes |
| 9 | FEC | Graph theoretic | $O(N + M)$ | no |
| 10 | BFC | Agglomerative | $O(V+E)$ | yes |

## 3. PERFORMANCE MEASURE

We have compared our BFC algorithm execution speed with Newman's Betweenness algorithm. We plotted a graph (shown in Fig. 3) between execution speed and size of network. The result shows the linear execution time complexity of BFC algorithm.

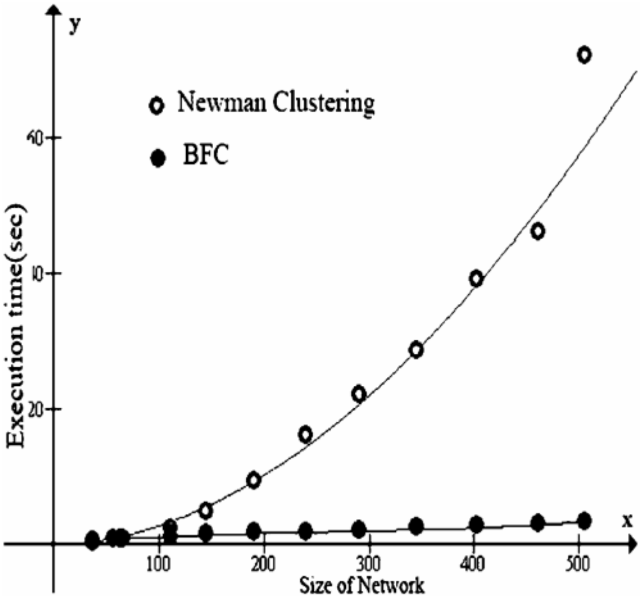


Fig. 3: Comparison with Newman's Algorithm

## 4. PROPOSED WORK

To mine out communities from a Social Network we have presented a new algorithm which we have called as BFC (Breadth First Clustering) because it proceeds in breath first way to traverse and incrementally extract out communities out of large Social Networks. We have implemented the algorithm on GUESS (Graph Exploration System) tool and have given the performance measure by comparing it with existing Newman's "Betweenness Algorithm" to show its effectiveness and linear time complexity.

## 5. CONCLUSIONS

We have presented a new algorithm BFC (breadth first clustering) which uses statistical approach for community mining in Social networks. The algorithm proceeds in breadth first way covering breadth of community and incrementally detects them from the Network. This algorithm is simple, fast and can be scaled for large Social networks. The effectiveness of this approach has been validated using network examples.

The time complexity of the algorithm is $O(V + E)$ where V represent number of nodes and E represent no. of edges in the network. The algorithm doesn't need any parameter to be supplied for its operation like cluster size or number (k) as in many other algorithms and It doesn't encompass complex iterative calculation of measures as in cut based approaches. So far we have tested this algorithm with medium sized networks, in the future we will enhance the algorithm to deal with large and dynamic networks of order higher than $10^5$.

## REFERENCES

[1] Bo Yang, W.K. Cheung, and Jiming Liu, "Community Mining from Signed Social Networks", IEEE / KDE, 19, No. 10, 2007.

[2] Bo Yang, Jiming Liu, "An Autonomy Oriented Computing (AOC) Approach to Distributed Network Community Mining", IEEE/ SASO 2007.

[3] Ying Zhou, Joseph Davis, "Discovering Web Communities in the Blogspace", 40th Hawaii International Conference, 2007.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.U. Hwang, "Complex Networks: Structure and Dynamics", Physics Reports 424, 175–308, 2006.

[5] Bo Yang, D.Y. Liu, "Force Based Incremental Algorithm for Mining Community Structure in Dynamic Network", J. Comp. Sci. & Tech., 21, No.3, May 2006.

[6] Rong Qian, Wei Zhang, Bingru Yang, "Detect Community Structure from the Enron. Email Corpus Based on Link Mining", ISDA, 2006.

[7] Ryutaro Ichise, Hideaki Takeda, "A Mining Method of Communities Keeping Tacit Knowledge", IEEE/ICDMW'06.

[8] Mohsen Jamali and Hassan Abolhassani, "Different Aspects of Social Network Analysis", IEEE/ WI'06.

[9] M. E. J. Newman, "Finding Community Structure in Networks using the Eigenvectors of Matrices", European Physics Journal, 0605087v3, 2006.

[10] Deng Cai, Zheng Shao, "Mining Hidden Community in Heterogeneous Social Networks", LinkKDD'05, 2005 ACM.

[11] Eric D. Kelsic, "Understanding Complex Networks with Community-finding Algorithms", SURF 2005.

[12] P. Pons and M. Latapy, "Computing Communities in Large Networks Using Random Walks", Proc. 20th Int'l Symp. Computer and Information Sciences (ISCIS '05), pp. 284-293, 2005.

[13] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society, Nature", 435, No. 7043, pp. 814-818, 9 June 2005.

[14] D.H. Kim and H. Jeong, "Systematic Analysis of Group Identification in Stock Markets", Physical Rev. E, 72, 046133, 2005.

[15] Jordi Duch, Alex Arenas, "Community Detection in Complex Networks using Extremal Optimization", Cond-mat/0501368 v1, 2005.

[16] M. E. J. Newman and M. Girvan, "Fast Algorithm for Detecting Community Structure in Networks", Physical Review E-69:026113, 2004.

[17] A. Clauset, M. E. J Newman & C. Moore, "Finding Community Structure in Very Large Networks", Physical Review E 70(066111), 2004.

[18] Andreas Noack, "An Energy Model for Visual Graph Clustering", GD 2003, Pages 425.436, Springer-Verlag, 2004.

[19] Gary William Flake, Robert E. Tarjan, and Kostas Tsioutsiouliklis, "Graph Clustering and Minimum Cut Trees", Internet Mathematics, 1, No. 4, 385-408, 2004.

[20] F. Radicchi, C. Castellano, "Defining and Identifying Communities in Networks", PNAS, 101, No. 9, pp. 2658-2663, 2004.

[21] M. E. J. Newman, "The Structure and Function of Complex Networks", SIAM Review, 45, 167–256, 2003.

[22] J.R. Tyler, D.M. Wilkinson, and B.A. Huberman, "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations", C&T '03, pp. 81-96, 2003.

[23] M. Girvan and M.E.J. Newman, "Community Structure in Social and Biological Networks", PNAS, 99, No. 12, pp. 7821-7826, 2002.

[24] G. W. Flake, S. Lawrence, C. Lee Giles, "Efficient Identification of Web Communities", ACM SIGKDD, 2000.

[25] S. Mancoridis, B. S. Mitchell, C. Rorres, Y. Chen, and E. R. Gansner, "Using Automatic Clustering to Produce High-level System Organizations of Source Code", In Proc. 6th IEEE International Workshop on Program Comprehension (IWPC 1998), Pages 45.52, IEEE, 1998.

[26] Thomas H. Cormen, Charles E. Leiserson and Ronald L. Rivest, Introduction to Algorithms, MIT Press, 1990.