IJCTA, 9(21), 2016, pp. 203-211

Community Detection Methods and Approaches in Social Networks-An Overview

U. Samson Ebenezar* N.P. Gopalan**

Abstract **:** Community detection in social network is a well known problem in computer science discipline and is of very much important to understand the structure and functioning of the networks. Nowadays social media like facebook, tiwtter, twoo, instagram, watsapp and flickr etc.. plays a vital role to bring all sorts of people to connect each other for sharing information in any format. Since the usage of social media increasing in exponential order and is the main source of information for entertainment, knowledge sharing, business and making relationship. Perhaps in any social network, the users are grouped under some communities. Researchers in social networks analysis have worked out many algorithms and methods for detecting the communities available in the network. In this paper we have enumerated few of community detection algorithms and methods so for discussed by the researchers over disjoint and overlapped communities in the social networks analysis.

Keywords : Social network, Graph, Clique, Partition, Community, Cut, MakeFuzzy, DOCNet, CESNA, CODICIL.

1. INTRODUCTION

Graph is a simplest and easily understandable tool to represent the networks such as computer, biological and molecular networks. In Social network analysis, graph theory is used for denoting the connections between the actors. In graph, the actors (nodes) are denoted as vertices and the connections between actors as edges. In computer networks, particularly in social networks detecting communities is a task of particular importance to the computing society. However, various approaches, methods and algorithms have been dealt over this topic for the past couple of decades.

Fig. 1. (*a*) Disjoint community **(*b*) overlapped community**

Community structure is a prevailing characteristic of social networks. The actors in the social network can be grouped into disjoint communities or overlapping communities [31]. Figure 1a, shows that only the vertices with in a set having edges is disjoint set, if there are some edges connecting vertices between the two or more sets then it is called overlapped sets as shown in figure 1b. If a network contains groups then communities exists in the

* Department of Information Technology Periyar Maniammai University, Thanjavur, Tamilnadu, India Email- u.samson@gmail.com

** Department of Computer Applications National Institute of Technology, Trichirappalli, Tamilnadu, India Email- npgopalan@nitt.edu

network [1]. It is hard to determine number communities in social network. So, the detection of communities in social networks is a non deterministic polynomial (NP) hard problem. The methods based on greedy algorithm gives poor result on highly complex social networks and these greedy algorithms for community detection needs the number of communities present in the network in advance, but it is hard to find in social networks [2], [27].

This paper is organized as follows, section II discusses the basic terminologies used in social networks, section III describes basic measure of social network analysis, section IV describing the traditional community detection methods, section V briefly explains the recent community detection methods proposed in various articles and section VI contains the conclusion and feature work.

2. BASIC TERMINOLOGIES

2.1. Social network

A Social network can be defined as a platform for creating social die among individuals in this computer era. A social network will be represented by a social graph G. The graph G contains a set of nodes. The total number nodes in the network are denoted as N. The nodes may be called as actors in social network. The edges in the graph represent the connection between two nodes, in graph, the connection between two actors x and y are depicted by the edge E_{xy}. The connections between the actors in the network can be depicted with the help of directed or undirected graph. For mathematical representation an adjacency matrix can be used. Let M be an adjacency matrix, an edge connecting the actors x and y will be taking the value $M_{xy} = 1$ otherwise $M_{xy} = 0$. Since the social networks do not fit into the topology of computer network, so it can be considered as a complex networks [3-4].

2.2. Community

In Social networks, various definitions are possible for a community. Thus the community or clusters in social network can be defined as groups of vertices which shares common properties or interest and do similar activity within the network [2]. Definition In [5] states that a community is a group of vertices in the network; inside the community most of the vertices are having connection with each other, but between the communities very few vertices are having connection with each other, shown in figure1(*b*) is example of overlapping communities. If no vertices in a community having connections with vertices in other community then the community is called disjoint community, as shown in figure 1(*a*).

2.3. Clique

Consider a graph G = (V, E), a clique in G is a subset of the vertices in $q \subseteq V$, that means any two individual vertices in a clique are adjacent to each other [7], the clique is otherwise called as maximal sub graph. In [6] a clique is defined as a maximal complete sub graph of a given graph, which is a group in a social network where every participant connected everyone else. Perhaps, identification of clique in social networks can be used for detecting the implicit community present in the network.

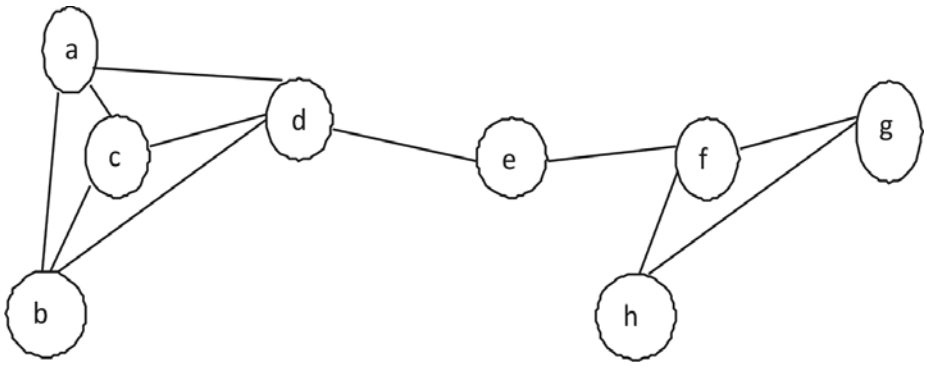

Fig. 2. Depicts a social network, It contains two cliques, First clique is formed by nodes {*a, b, c, d*}, Second clique is formed by nodes {*f, g, h*}.

From the figure 2 it can be concluded that identification of clique can be used for detecting both disjoint and overlapped communities. The various forms of cliques such as *n*-clique, *n*-clan, *n*-club and *k*-plexes are clearly explained in [2].

The edges running between the groups are called cut. If the cut is removed the graph is divided in to disjoint partitions. If the removal of edge is in single size then it is called minimal cut size and the removal of edges more than two edges from the graph then it is as called maximum cutsize. The Figure 2 shows difference between the minimum and maximum cut size. The edges connecting the vertices *d*,*e*, edge (*de*) and *e*,*f*, edge(*ef*) are minimum cutsize, because if one remove any one of the edges from the graph, it will be split in to two sets. Then the edges (*ad*), (*cd*), (*bd*) are called maximal cutsize, because here more than two edges are involved.

3. BASIC MEASURES IN SOCIAL NETWORK ANALYSIS FOR COMMUNITY DETECTION

Here, the different centrality measures used for detecting communities in the social network is listed.

3.1. Degree centrality ($DC_D(n)$)

[30] is a measure for finding the central actor or participant who is influencing the network. If any vertex or node having highest degree centrality, that node will be the influencing node in the community structure. The degree centrality can be calculated as [8]

$$DC_D(n_i) = \frac{d(n_i)}{y-1} \quad (1)$$

here, $d(n_i)$ denotes degree of every node n_i and **y** –1 denotes the number of nodes excluding the central node. If any node has highest value for $C_D(n_i)$ that node will be considered as central node.

3.2. Closeness centrality($C_c(n)$)

[30] is useful when to calculate the reachable distance between two nodes or actors, as explained in [8], Closeness centrality of actors p and q can be calculated as [8]

$$C_c(n_i) = \frac{y-1}{[\sum_{q=1}^{y} d(n_p, n_q)]} \quad (2)$$

Here, *y* is total nodes in the network. Closeness is a measure of shortest distance between two nodes along the connection path to pass the information.

3.3. Betweenness centrality($BC_b(n)$)

[30] is a measure for an actor or node to what extent it has control over the information flow between nodes or actors in the network. It is very much useful when one need to find the connections between the overlapping communities [8, 10].

$BC_b(n)$ of a node n is calculated as [9] $BC_b(n) = \sum_{x \neq n \neq t} (\beta_{xy}(n)/\beta_{xy})$ (3)

where *x* and *t* are two nodes in the network different from the node *n*, β_{xy} is total shortest paths from *x* to *t*, and $\beta_{xy}(n)$ is the number of shortest paths from *x* to *t* that n lies on.

3.4. Vertex similarity

[32] is the measure to find the similar nodes in the neighborhood for clustering; using this one can identify an implicit community in the network. Vertex similarity is calculated based on jaccard similarity and cosine similarity, the calculation and formulas are clearly given in [21].

3.5. Normalized Mutual Information

(NMI) is a measure used for validating the cluster quality to detect the community. However, in reality number of communities formed after grouping the nodes in the network will be different from the predicted value before clustering. Moreover, the clustering result is different from the ground truth.

Normalized Mutual Information (NMI) can be calculated [28] as follows

$$\text{NMI}(X ; Y) = \frac{\text{MI}(X ; Y)}{\sqrt{\text{K}(X)\ \text{K}(Y)}}$$

MI (X, Y) is mutual information shard among two distributions and K (X) entropy of distribution X and K (Y) entropy of distribution Y. Normalized Mutual information and entropy methods are clearly explained in [28].

3.6. Omega index

[28] is the modified version of the Adjusted Rand Index (ARI), The Omega index is a measure for considering how many pairs of nodes are exactly placed together in no clusters, one clusters, two clusters, and so on in the network. When the omega index is used the network is first divided in large partitions are called covers, each cover may have many communities. This index is based on the agreement between the node pairs in two covers. If two nodes are having agreement means they should be placed in the same community.

The Omega index of covers Co1 and Co2 can be calculated as [28]

$$\text{Omega Index} - \omega(\text{Co1}, \text{Co2}) = \frac{\omega_u\ (\text{Co1}, \text{Co2}) - \omega_n\ (\text{Co1}, \text{Co2})}{1 - \omega_n\ (\text{Co1}, \text{Co2})} \quad (5)$$

Here ω_u (Co1, Co2) is unadjusted omega index and ω_n (Co1, Co2) is null model omega index for covers Co1 and Co2.

4. TRADITIONAL COMMUNITY DETECTION METHODS AND APPROACHES

4.1. Graph partitioning method

The Graph is divided in to numbers of groups of predefined size then it is called graph partitioning. Between the groups, the number of edges connecting the vertices should be less [2].The number of edges connecting the groups is called as cut size. If the cut is removed, the graph is partitioned into two disjoint groups. The vertices in each group are having similar property. If the graph partitioning method is used for dividing the graph, the number groups and the size of groups must be given in advance. Otherwise the partitioning method will not be suitable for detecting the communities. This method, grouping together the minimum degree vertex in to one cluster and the remaining vertices grouped into another cluster. The removal of two or more edges will partition the graph in to multiple clusters then the number of edges is removed is called multi cut.

4.2. Hierarchical clustering method

Clustering is the process of partitioning the graph in to sub groups is called clusters. However, the number of clusters will be formed and the size of each clusters hardly known in advance. In that situation graph partitioning method may not be helpful, if the number of clusters and cluster size is assumed in advance that decision may end up with errors. Often, the graph exhibits hierarchical structure property, which displays several levels of clusters. In that smaller clusters found in larger clusters, which are again included in large clusters in the network, this process will be repeated until all clusters arranged in a hierarchical manner. To exhibit the multilevel arrangement of the graph [2-3], this method can be used. Hierarchical clustering is found in social, biological, engineering, and marketing networks.

In hierarchical clustering, similarity measure between vertices must be defined in the initial stage. The similarity measure is used for computing the similarity between every pair of vertices in the network, no need to consider whether they are connected with each other or not. After this computation is over, a new $n \times n$ similarity matrix X will be formed. The vertices having high similarity only will be grouped in this technique. It is based on two algorithms as follows,

Agglomerative algorithms : In this high similarity vertices are merged with a cluster iteratively. It is following the bottom up approach, as it starts from a cluster with single vertices and at the end it construct a graph with a unique cluster contains hierarchy of vertices.

Divisive algorithms : It splits the cluster iteratively by the removal of edges lying between the vertices having low similarity measure. Newman-Girvan algorithm described in [9] is based on the divisive algorithm.

Number of clusters and size of the clusters do not require in advance in hierarchical clustering method. However, it does not provide a way to discriminate between the partitions obtained by the procedure, and to choose the cluster or clusters that better represent the community structure of the graph.

4.3. Clique percolation method

CPM methods assume communities are constructed by multiple adjacent cliques. Based on the original approach [14], the Sequential Clique Percolation (SCP) [13] algorithm sequentially generates cliques to form connected communities. Another kind of the approaches maintains a tree, which is a multilevel structure reorganized from the original graph, aiming at ûnding communities corresponding to the branches of the tree [15].

4.4. Label propagation method

This method starts from local neighborhood to recognize communities automatically. The Label propagation algorithm (LPA) [16] adopts an asynchronous update strategy where nodes join in groups under their neighbors' choices. The HANP algorithm [17] based on Hop Attenuation and Node Preference adopts additional rules to ensure more stable and robust results [15].

5. RECENT COMMUNITY DETECTION METHODS AND APPROACHES

5.1. DOCNet (Detection of Overlapping Communities in Networks)

It is an efficient community identification approach for overlapping communities [12]. The DOCNet (Detection of Overlapping Communities in Networks), contains the prerequisite to identify the overlapping communities in social networks. This model is based on agglomerative hierarchical clustering approach, since communities are built in collective manner. This algorithm starting from a single node and repeatedly expands its border nodes until it reaches an equilibrium state. This method uses the index connectivity as the objective function and node importance (NI) and membership degree of each node as metrics [12].This algorithm is suitable for finding the overlapped communities in the network. Thus, DOCNet consists of two main components:

5.1.1. Building the core of a community

DOCNet begin with initially empty partition. As a starting point, consider each vertex as a community C. Then node importance (NI) of all nodes of the graph is computed. Then it sorts vertices according to their importance and in descending order in the vector Imp. These steps represent the initialization phase. The next core of the community will be formed. First, the most important node from vector Imp is selected, this node is the ''most inûuential'' in the remaining non-partitioned part of graph. Next, build the ''core of community'' of C formed by its center and its border.

5.1.2. Extending its core.

The core of the community is extended based on largest membership degree to the core of the community. [12] The set of nodes situated on the border of C is denoted by K_C and which are candidates to its extension. Regarding the extension stage of C, the algorithm proceeds as follows; It chooses the candidate node n_c from K_C with the largest membership degree to the core community. Then it starts by adding this node. If it increases the objective function then it update all boundaries nodes and their membership degrees, and it checks again the next vertex in K_C. Otherwise, it stops the expansion of this community, and remove community members from Imp and move to the formation of the next community.

5.2. The Cores-Aware and the Cores-Unaware algorithms

The core-aware and cores unaware algorithms proposed in [5], contain keys and peaks as measures. If the degree of a vertex is no less than any of its neighbor, then the vertex is called as Key. And if a vertex has the largest

degree in a community, then it is called as Peak. Peaks and Keys are deûned as follows. Keys is a set of Vertices with in a network. Each vertex in keys is called key. Peaks is a set of vertices within a community. Each Vertex in Peaks is called Peak.

5.2.1. Cores-Aware algorithm

In some cases cores of the community is known in advance. For example, if the goal is to ûnd communities in a social network for an election, the candidates are considered as cores [5]. For community detection first the Cores-Aware algorithm is used then the Cores unaware algorithm followed when the Cores are not known.

This algorithm follows breadth- first traversal method. The core-aware algorithm is as follows.

1. In a network, first the cores are checked and assigned a unique label to every core, and in the initial stage count of cores and distance between the cores are initialized as 1.
2. The graph follows parallel breadth- First traversal. As explained in [5], consider a vertex v and sometimes any one of the neighbors may not be checked. Then the vertex v will be checked whether it has a core with the unchecked neighbor. If a vertex has unchecked neighbor that unchecked neighbor will taken as a candidate. Then all neighbors of the candidate will be updated with information v and it's the count is set to 1. If two vertices commonly shares a candidate clusters, the count entries are increased by 1 for all shared candidate clusters. Then other neighbors are checked with the similar method.
3. The minimum distance vertex from its core will be checked and added into the community. If two communities are having same distance from their cores, the community having larger number of edges will be selected.

5.2.2. Cores-Unaware algorithm

The cores of the communities are not known in advance in many cases. To handle such instances, the Cores-Unaware algorithm was introduced, which does not need the information of cores. The algorithm works as follows; it is common that some communities do not have certain cores. If some candidates can be selected as cores, then community detection can be processed by invoking the Core-Aware algorithm. Here the problem is which vertices should be selected as the Cores. To solve this problem, this algorithm chooses candidates of cores and then ûlters the candidates with structural constraints of the network. Since the incorrect cores may mislead the community detection so the cores should be chosen carefully. Here the vertex with the largest degree must be belonging to the core. These algorithms are efficient and can take advantage of feedback and structural information to improve the performance. Based on the conclusion in the paper [5], it has linear computational complexity and suitable for large network [5].

5.3. Communities from Edge Structure and Node Attributes (CESNA)

CESNA (Communities from Edge Structure and Node Attributes) was introduced in [18], is an efficient method for detecting overlapping communities based on the attributes shared by the actors or participant in the social network. Because many people in social network communication share similar attributes, so it will be suitable for detecting hidden communities.

CESNA is a probabilistic model, considers the social network elements such as memberships of the community, topology of the network and node attributes. The working principle of the model is as follows;

1. Nodes in the communities will be connected to each other if the nodes are belonging to the same community.
2. Overlapping will be possible among the communities, and individual nodes may be belonging to multiple communities.
3. Nodes belonging to more than one common community will be easily connected each other than if the nodes are belonging to common community (*i.e.*, overlapping communities are denser [19], [20]).
4. Mostly all of the nodes share common attributes if they belong to the same community. A community may contain members from same college or work place can be considered as examples.

The process of CESNA as follows; It assumes a network P, contains N nodes and C communities each node has K attributes. The network is denoted by P, the node attributes denoted by Y (Y_{vk} is *k*-th attribute of node *v*), and community memberships by F. If node belong to the community C than it should have non- negative community membership affiliation weight $F_{vc} \in [0, \infty]$ for the community membership F. If affiliation weight $F_{vc} = 0$ than the node *v* does not belong to community *c*. Incorporating node attributes into community detection gives two direct advantages. One is the improved accuracy in community detection, and the second advantage is that the node attributes useful for interpreting detected communities.

5.4. Community Discovery Inferred from Content Information and Link-structure (CODICIL)

The method proposed in [21] is based on content and link information. This algorithm identifies the important edges in the network based on content information of nodes and topology of the network to retain the edges in the community. These methods rely on combining content and topological or link information in a usual way. CODICIL creates content edges based on content similarity among the neighbor nodes, the content similarity is computed using cosine similarity method [21]. Then it retains the edges which are relevant among the neighborhood nodes, this process continues until a simplified graph is formed based on union of content and topological edges. The content and topological edges are identified based on the information shared by nodes through the edges. At the end the clusters will be formed in the graph using the content –insensitive clustering algorithm such as METIS or Markov clustering.

5.5. Fuzzy based community detection method

The MakeFuzzy techniques introduced in [22] is a fuzzy based community detection algorithms for overlap communities. Many algorithms namely CFinder [14], CONGA [23], LFM [24], and COPRA [23] are "crisp" in nature. Here, in this method Fuzzy Rand Index is used as a metric to partition the network. The fuzzy and crisp networks differ in two respects. In crisp network the fraction of fuzzy rand index value is less than 1 for overlapping vertices, the expected degree of two overlapping community vertex is greater than that of a non overlapping community vertex; but there is a variation in fuzzy networks. Next, in crisp networks when all vertices are overlapping, each vertex equally belongs to its two communities, but the same not necessary in fuzzy networks. If by mistake a vertex is assigned to a single community by any algorithm can get a higher score on a fuzzy network than on a crisp network.

6. CONCLUSION

Community detection in social network is the well known problem and has been discussed for couple of decades; it is the importance of research community in social network. Because in today's world social networks playing vital role to connect people of various domain from any part of this world. This community detection has many potential applications namely trend analysis in citation network, improving the capability of recommender system to give accurate recommendation and evaluation of communities in social media, etc**.** The finding and investigation of communities in social network is useful for commercial, educational and developmental purposes. In this paper we discussed some of the concepts of traditional methods in community detection namely graph partition, hierarchical clustering, clique percolation and label propagation methods. Then we discussed some the recent methods such as DOCNet, CESNA, CODICIL, Core-aware and Core-unaware and the fuzzy based community detection methods. Our research work will be focused on constructing an algorithm for community detection based on fuzzy inference system.

7. REFERENCES

1. Özturk K. Community detection in social networks. Msc. Thesis. Graduate School of Natural and Applied Sciences, Middle East Technical University, 2014.
2. Fortunato S. Community detection in graphs. Phys Rep 2010, 486:75–174.doi:10.1016/j. physrep.2009.11.002.
3. Fasmer EE. Community detection in social networks. Master Thesis. Department of Informatics, University of Bergen, 2015.
4. Barabási A-L, Albert R. Emergence of scaling in random networks. Science 1999, 286:509–512.doi:10.1126/ science.286.5439.509.
5. Yakun Li, Hongzhi Wang, Jianzhong Li, Hong Gao ,Efficient community detection with additive constrains on large networks. Elsevier- Knowledge-Based Systems -52(2013) 268-278.
6. https://www.safaribooksonline.com/library/view/social-network analysis/9781449311377/ch04.html
7. https://en.wikipedia.org/wiki/Clique_(graph_theory)
8. M. E. J. Newman , A measure of betweenness centrality based on random walks, Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109–1120
9. M. Girvan and M. Newman, Community Structure in Social and Biological Networks, Proceedings of the National Academy of Sciences, vol. 99, no. 12, pp. 7821– 7826, Jun. 2002.
10. http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/
11. Lei Tang , Huan Liu, Community Detection and Mining in Social Media, Synthesis Lecture on Data Ming and Knowledge Discovery, Morgan & claypool Publishers, USA.
12. Delel Rhouma , Lotû Ben Romdhane , An efficient algorithm for community mining with overlap in social networks, Elsevier- Expert Systems with Applications 41 (2014) 4309–4321,
13. J. M. Kumpula, M. Kivel¨a, K. Kaski, and J. Saram¨aki. Sequential algorithm for fast clique percolation. Physical Review E, 78(2):026109, 2008.
14. G. Palla, I. Der´enyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435(7043):814–818, 2005.
15. Meng Wang , Chaokun Wang , Jeffrey Xu Yu , Jun Zhang , Community Detection in Social Networks:An In-depth Benchmarking Study with a Procedure-Oriented Framework, Proceedings of the VLDB Endowment, Vol. 8, No. 10.
16. U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 76(3):036106, 2007.
17. I. X. Leung, P. Hui, P. Lio, and J. Crowcroft. Towards real-time community detection in large networks. Physical Review E, 79(6):066107, 2009.
18. Jaewon Yang, Julian McAuley, Jure Leskovec, Community Detection in Networks with Node Attributes,Stanford University,USA.
19. S. L. Feld. The focused organization of social ties. American J. of Sociology, 1981.
20. J. Yang and J. Leskovec. Structure and overlaps of communities in networks. ACM TIST, 2013.
21. Yiye Ruan, David Fuhry, Srinivasan Parthasarathy, Efûcient Community Detection in Large Networks using Content and Links, The Ohio State University, USA.
22. Steve Gregory , Fuzzy overlapping communities in networks, University of Bristol, Bristol BS8 1UB, England.
23. Gregory S 2010 Finding overlapping communities in networks by label propagation New J. Phys. 12 103018
24. Lancichinetti A, Fortunato S and Kertész J 2009 Detecting the overlapping and hierarchical community structure of complex networks New J. Phys. 11 033015
25. Guanbo Jia , Zixing Cai , Mirco Musolesi , Yong Wang , Dan . Tennant , Ralf. J. M. Weber , John K. Heath , and Shan He, Community Detection in Social and Biological Networks using Differential Evolution, Learning and Intelligent OptimizatioN Conference LION 6, Paris, Jan 16-20, 2012

26. Collins, L. M. And Dent, C. W. 1988. Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. Multivar. Behav. Res. 23, 2, 231–242.

27. Punam Bedi and Chhavi Sharma, Community detection in social networks, WIREs Data Mining Knowl Discov 2016, 6:115–135. doi: 10.1002/widm.1178.

28. Xie, J., Kelley, S., and Szymanski, B. K. 2013. Overlapping community detection in networks: The state-of- the-art and comparative study. ACM Comput. Surv. 45, 4, Article 43 (August 2013).

29. Neveen Ghali , Mrutyunjaya Panda, Aboul Ella Hassanien, Ajith Abraham, Vaclav Snasel Social Network Analysis : Tools, Measures and Visualization , Springer Link, Chapter 1, PP 3-23. 14 june 2012.

30. Joshua D. Guzman, Richard F.Deckro, Mathew J. Robbins, James F. Morris and Nicholas A Ballester , An Analytical Comparison of Social Network Measures, IEEE Transactions on Computational Social Systems, Vol. 1 No.1, March 2014.

31. M.E.J. Newman, Detecting community structure in networks, Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109–1120, The European Physical Journal B.

32. Hung-Hsuan Chen , Liang Gou , Xiaolong (Luke) Zhang, C. Lee Giles, Discovering Missing Links in Networks Using Vertex Similarity Measures, ACM 978-1-4503-0857-1/12/03, SAC'12 March 25-29, 2012.