Disjoint Community Detection Algorithms in Social Media: An Overview

K Swapnika
Computer Science Engineering Department, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, INDIA

K Swanthana
Computer Science Engineering Department, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, INDIA

Dr Y VijayaLatha
Head, Information technology, Department, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, INDIA

Abstract—Communities are defined as partition of the set of vertices, that each node is put into one and only one community. In some cases vertices are present in more than one community. This might happen in a social media network where each vertex represents a person, and the communities represent the different groups of friends. The variety of methods that have appeared for detecting communities is even larger, because for each community definition there is more than one method claiming to detect the respective communities. To analyze the link structure of a sample social media, the link structure is represented as a graph, associating several attributes with the vertices and edges. Each vertex represents a community, and each edge represents a link. The investigation of the community structure in the social media has been an important issue in many domains and disciplines. Community structure has more significance with the increasing popularity of online social network services like Facebook, MySpace, or Twitter. We mainly discuss about various community detection algorithms in real world networks. This paper represents an overview of the community detection algorithms in social media.

Index Terms— algorithms, community, graph, node, social media.

I. Introduction

The proposed survey discusses the topic of community detection and its algorithms in the context of Social Media. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges join vertices of different clusters. Furthermore, the distribution of edges is not only globally, but also locally inhomogeneous, with high concentrations of edges within special groups of vertices, and low concentrations between these groups. This feature of real networks is called community structure [21], or clustering.

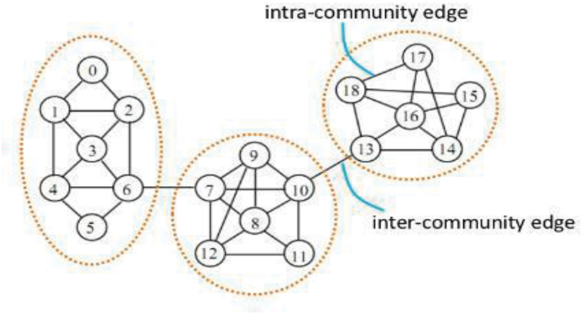

Fig 1: A simple graph with three different communities

In the above figure, we can see there are three communities, in which all the nodes contain in a community are dense and intra-connected with each other. The nodes in a community are interconnected sparsely to the other community of nodes. In a community, nodes are connected with each other based on their human relationship like friendship, colleague etc.

There is no unique and widely accepted definition of community. Community definitions are formulated with reference to the network structure of the system under study and are commonly bound to some property either of some set of vertices (local definitions) or of the whole network (global definitions). Social Media communities can be further described as *explicit* or *implicit*. Explicit communities are created as a result of human decision and acquire members based on human consent. Examples of explicit Social Media communities are Facebook and Flickr Groups. Implicit communities, on the other hand, are assumed to exist in the system and "wait" to be discovered. Implicit communities are particularly important for two reasons [28]: (a) they do not require human effort and attention for their creation and (b) they enable the study of emerging phenomena within Social Media systems.

This survey focuses on the definition and discovery of the implicit communities.

A. Implicit community classes

Implicit communities are defined with reference to the network structure. The most established notion of community-ness within a network is based on the principle that some sets of vertices are more densely connected to each other than to the rest of the network. Depending on whether this property of vertices is considered locally (on a connected subset of vertices) or globally (on the whole network), we distinguish between local and global community definitions. Communities are also defined on the basis of the result of some principled network-based process also called as process-based definitions. These definitions are described in the paper by Kovacs [17, 28].

B. Local Definition

The local definition focuses on the concepts of subgroup cohesiveness and mutuality. Examples of such community definitions are *cliques, n-cliques, n-clubs, n-clans, k-plexes, and k*-cores [4]. However, the above definition is too restrictive and computationally very expensive. Thus, their use is very limited in a Social Media context. Alternatively, the internal vertex, external vertex and sub-graph degrees have been used to define community-ness. The internal degree of a vertex is the number of edges that connect it to vertices of the same community. The external degree is defined in a similar way. The definitions of communities in the *strong* and *weak* sense [24] are based on the internal and external degrees of vertices belonging to a community.

C. Global Definition

Global community definitions consider community structure as a whole network. There are several important classes of global community definitions. The most direct community definition is based on the number of edges falling between the communities (cut size) as a measure of quality of a given network partition into communities. Since the absolute number of inter-community edges is problematic, normalized measures such as *Normalized Cut* [27] and *conductance* [14] have been introduced for quantifying the profoundness of separation between the communities of a network. A wide class of global community definitions relies on some similarity measure between network vertices. Once pairwise similarities between vertices are computed, communities are defined as clusters of vertices that are close to each other.

D. Process-Based Definition

In this, defining community is done by considering some community formation process on the network. For instance, the Clique Percolation Method [22] considers a *k*-clique template that "rolls" on the network and results in a community consisting of the union of all *k*-cliques that are adjacent to each other (i.e. share *(k*−1*)* nodes). This dynamic process is iteratively applied on the network in order to reveal groups of vertices that form well-separated communities. Another dynamic process used for the definition of communities is the synchronization of a set of phase oscillators on a network: groups of vertices whose oscillators synchronize first are considered to form the communities. Finally, a label propagation scheme based on neighbor majority voting is devised by Raghavan [25] to define communities as groups of vertices forming stable consensus with respect to their label.

II. TYPES OF COMMUNITIES

Communities can be of two types:

A. Disjoint community

Disjoint community is also known as crisp assignment,
where the relationship is between a node and a community. Here a node belongs to single community.

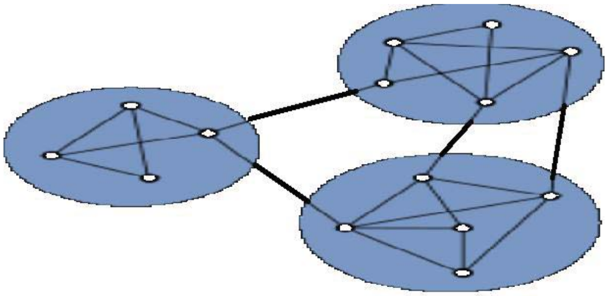

Fig 2: Disjoint communities

*B. Overlapping community***:** Here a node may belong to more than one community. This is known as fuzzy assignment of nodes, where a node may belong to more than one community.

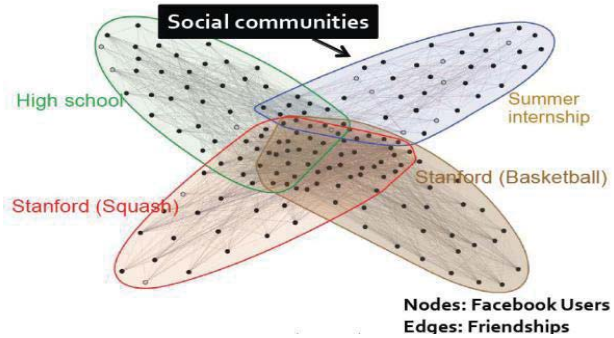

Fig 3: Overlapping communities in Facebook network

III. DISJOINT COMMUNITY DETECTION ALGORITHMS

Community detection methods

The variety of methods that have appeared in this literature for detecting communities is even larger, since for each community definition there is more than one method claiming to detect the respective communities. In this section, we will summarize the most important classes of such methods, associate them with the definitions presented in the previous section and comparatively discuss their performance requirements in terms of computational complexity and memory consumption, as well as their dynamic computation characteristics, which are particularly pertinent for the analysis of Social Media networks.

The survey classifies community detection algorithms into below categories [7]:

A. Traditional
B. Divisive
C. Model-based
D. Dynamic Algorithms

A. *Traditional Methods*

There are three types of traditional methods to detect communities in a social network.

1. *Graph Partitioning*

Graph partitioning method represents to divide the nodes in g groups of predefined size, such that the number of edges lying between the groups is minimal. The number of edges running between clusters is called cut size [7]. Below figure presents the solution of the problem for a graph with six nodes, for g = 2 and clusters of equal size. Many algorithms perform a bisection of the graph. Generally, partitions into more than two groups are achieved by iterative bi-sectioning.

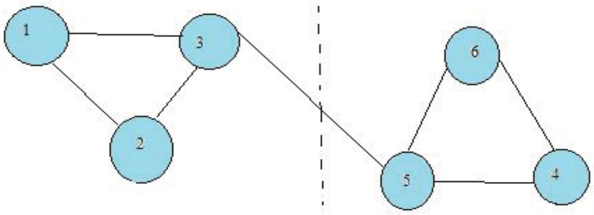

Fig 4*:* graph partitioning method. The green dashed line shows the solution of the minimum bisection problem for the graph illustrated.

2. *Hierarchical clustering*

Hierarchical clustering is a widely used data analysis tool. The idea behind this clustering is to build a binary tree of data that merges similar groups of points. It is not easy to know the number of clusters if the graph is split. if the graph is in hierarchical structure with small groups included within larger groups, in that case hierarchical clustering algorithm may be used.

3. *Spectral clustering*

Donath and Hoffmann [8] contributed first on spectral clustering in 1973, they used eigen vectors of the adjacency matrix to partition the graph. Spectral clustering makes use of eigen values of the similarity matrix of the data. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset. Andrew Y .Ng [4] have analysed the algorithm of spectral clustering as the ideal case and the general case.

B. Division algorithms

In hierarchy clustering methods, such as Radicchi [10] and Spectral [18], gradually separate the entire network into local parts by the edge clustering coefficient or the eigenvalue of modularity matrix. Direct partitioning methods separate the entire network into disjoint communities. The Scalable Community Detection (S.C.D) algorithm [1] partitions the network by maximizing the weighted community clustering [2], a recently proposed metric of community. Maximal k-Mutual-Friends (M-K.M.F) [11] algorithm incrementally filters out the connections by the number of mutual friends between nodes to let the communities spontaneously emerge. Conversely, the second kind of approaches takes a bottom-up manner from local structures to the whole graph, and the communities are formed during this process. Label propagation methods start from local neighborhood to recognize communities automatically.

1. *Girvan–Newman algorithm*

According to survey, this algorithm belongs to the category of divisive algorithms [19]. Its underlying principle calls for removing the edges that connect different communities. In the algorithm described, several measures of edge centrality are computed, in particular the so-called intermediate centrality, whereby edges are selected by estimating the level of edge importance based on these measures. As an illustration, intermediate centrality is defined as the number of shortest paths using the edge under analysis. The steps involved are as follows:

1. Compute centrality for all edges,
2. Remove edges with the greatest centrality (when ties exist with other edges, one edge is to be chosen at random),
3. Recalculate centralities on the remaining graph,
4. Iterate beginning at step 2.

This work has exerted great influence on research and, consequently, edge centrality has been a key field of study for many scientists, resulting in the proposal of several measures.

C. Model-Based Methods

High values of modularity represent good partitions of a graph. There are four techniques discussed as below

1. *Greedy Techniques*

Newman [20] proposed first greedy method to maximise modularity. It is a hierarchical clustering method where edges do not contain in the graph initially; edges are added one by one during the procedure.

2. *Simulated Annealing*

Simulated Annealing [16] is probabilistic procedure used in different fields and problems. This procedure consists of the space of possible states looking for the maximum global optimum of a function F. Guimera [12] first applied simulated annealing for modularity optimization. The standard implementation [13] of them combines two types of moves: local moves, where a single node is shifted from one cluster to another randomly; and global moves, which consist of mergers and splits of communities.

3. *Extremal Optimization*

Boettcher and Percus [6] proposed Extremal optimization; In an interrogative search procedure. This technique is based on the optimization of local variables. Duch and Arenas [9] used this technique for modularity optimization. Modularity can be measured as a sum of the nodes in the graph. We can get a fitness measure for each node by dividing the local modularity of the node by its degree. Degree of the node does not define the measure.

4. *Spectral Optimization*

In this by using the eigen values and eigen vectors of a spectral matrix, modularity can be optimized. Wang [30] used community vectors to achieve high-modularity by partitioning into a number of communities smaller than a given maximum. If the eigen vectors are taken corresponding to the two largest eigenvalues, then we can obtain a split of the graph in three clusters.

5. *Spectral Algorithms*

Here we compute the transition probabilities by enabling the conductances for a random walker moving on the graph, and from the transition probabilities, we can build a similarity matrix between the node pairs. Hierarchical clustering is applied to join nodes in communities. If we need to compute the whole spectrum of the laplacian matrix, the time taken by this algorithm is $O(n^3)$, the algorithm proposed by alves [3] is slow.

D. *Dynamic Algorithms*

In this type there are three algorithms to be discussed: Spin models, Random walk, and Synchronization.

1. *Spin Models*

In statistical mechanism, the Potts model [31] is the most popular model. This model elaborates a system of spins that can be in q different states. It favors spin alignment such that all spins are in the same state at zero temperature. That means the interaction is ferromagnetic in this model. The ground state of the system may not be the one where all spins are aligned if antiferromagnetic interactions are also present. But, different spin values coexist in homogeneous clusters in a state. Based on Potts model, in 2004, Reichardt and Bornholdt [26] proposed a method to detect communities that maps the graph onto a zero-temperature q-Potts model with nearest-neighbor interactions.

2. *Random Walk*

In 1995, Hughes [15] showed that random walk can be useful to detect the clusters in a graph. If a graph contains several clusters, a random walker spends a long time inside a cluster due to the high intra-connections among all the nodes. All clustering algorithms based on the random walk can be trivially extended to the case of weighted graphs. In 2004, Zhou and Lipowsky [32] used biased random walkers, where the bias happens to the fact that walkers usually move towards the nodes sharing a large number of neighbors with the starting node in a graph. A proximity index is defined to show that how much a pair of nodes is closer to all other nodes in the graph. The procedure is called netwalk to detect the communities in a graph.

3. *Synchronization*

Synchronization [23] is an excellent process occurs in the systems and interacts among the units in nature and technology. All the units of the system are in the similar state at every moment while the system is in synchronized state. To detect the communities in a real world network, synchronization can also be applied. In 2007, Boccaletti [5] have designed a method for community detection applying the concept of synchronization.

IV. Conclusion

In this paper, we discussed the fundamental concepts of community in social media, then in the next section, we have elaborated the existing algorithms to detect the communities in social media. The paper briefly illustrated the traditional methods, divisive algorithm, modularity based methods, spectral algorithms and dynamic algorithms to detect the communities in real world networks. We hope the concepts demonstrated in this paper to detect the communities in real world networks will help us to study the community structure in social media deeply in future for big data community detection**.**

REFERENCES

[1] A. Prat-P´erez , D. Dominguez-Sal, and J.-L. Larriba-Pey. “High quality, scalable and parallel community detection for large real graphs”. WWW, pages 225–236, 2014.

[2] A. Prat-P´erez, D. Dominguez-Sal, J. M. Brunat, and J.-L. Larriba-Pey.
“Shaping communities out of triangles”. CIKM, pages 1677–1681, 2012.

[3] Alves, N. A., Phys. Rev. E 76(3), 036101, 2007.

[4] A.Y. Ng, M.I. Jordan, Y. Weiss, "On Spectral Clustering: Analysis and Algorithm", Stanford AI Lab.

[5] Boccaletti, S., M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda, Phys. Rev. E 75(4), 045102, 2007.

[6] Boettcher, S., and A. G. Percus, Phys. Rev. Lett. 86, 5211, 2001.

[7] Deepjyoti Choudhry, Arnab paul “Community Detection in social Networks: An Overview” volume 2, Special Issue:02, December 2013.

[8] Donath, W., and A. Hoffman, IBM Journal of Research and Development 17(5), 420, 1973.

[9] Duch, J., and A. Arenas, Phys. Rev. E 72(2), 027104, 2005.

[10] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi.
“Defining and identifying communities in networks”. PNAS,
101(9):2658–2663, 2004.

[11] F. Zhao and A. K. Tung. “Large scale cohesive subgraphs discovery for social network visual analysis”. VLDB, pages 85–96, 2012.

[12] Guimera, R., M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E 70(2), 025101 (R) , 2004.

[13] Guimera, R., and L. A. N. Amaral, Nature 433, 895, 2005.

[14] Kannan R, Vempala S, Vetta A (2004) “On clusterings: good, bad and spectral”. J ACM 51(3):497–515.

[15] Hughes, B. D., “Random Walks and Random Environments: Random Walks” Vol. 1, Clarendon Press, Oxford, UK, 1995.

[16] Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi, Science 220, 671, 1983.

[17] Kovács IA, Palotai R, Szalay MS, Csermely P (2010) “Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics”
PLoS ONE 5(9):e12528.

[18] M. E. Newman. “Modularity and community structure in networks”.
PNAS, 103(23):8577–8582, 2006.

[19] M. Girvan and M. Newman, “Community Structure in Social and Biological Networks”, Proceedings of the National Academy of Scinces, vol. 99, no. 12, pp. 7821– 7826, Jun. 2002.

[20] M.E.J. Newman and M Girman, “Finding and evaluation community structure in networks”, Physical Review E, 69(2), 2004.

[21] Mini Singh Ahuja, Jatinder Singh, Neha “Overlapping Community detection Algorithms:-A Review” Volume:02 Issue:09, December 2015.

[22] Palla G, Derenyi I, Farkas I, Vicsek T (2005) “Uncovering the overlapping community structure of complex networks in nature and society”. Nature 435(7043):814–818.

[23] Pikovsky, A., M. G. Rosenblum, and J. Kurths, “Synchronization : A Universal Concept in Nonlinear Sciences”, Cambridge University Press, Cambridge, UK, 2001.

[24] Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004)
“Defining and identifying communities in networks”. Proc Natl Acad
Sci USA 101:2658–2663.

[25] Raghavan UN, Albert R, Kumara S (2007) “Near linear time algorithm to detect community structures in large-scale networks”. Phys Rev E 76:036106.

[26] Reichardt, J., and S. Bornholdt, Phys. Rev. Lett. 93(21), 218701, 2004.

[27] Shi J, Malik J (2000) “Normalized cuts and image segmentation”.
IEEE Trans Pattern Anal Mach Intell 22(8):888–905.

[28] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, Ploutarchos Spyridonos “Community detection in Social Media Performance and application considerations” Springer Published online: 14 June 2011.

[29] Wasserman S, Faust K (1994) “Social network analysis: methods and applications”. Cambridge University Press, Cambridge.

[30] Wang, G., Y. Shen, and M. Ouyang, Comput. Math. Appl. 55(12), 2746, 2008.

[31] Wu, F. Y., Rev. Mod. Phys. 54(1), 235, 1982.

[32] Zhou, H., and R. Lipowsky, Lect. Notes Comp. Sci. 3038, 1062, 2004.