International Conference on Information and Communication Technology and Its Applications (ICTA 2016)
Federal University of Technology, Minna, Nigeria
November 28 – 30, 2016

Community Detection in Networks: Algorithms and Challenges

Elizabeth N. Onwuka, Bala A. Salihu, and Paschal S. Iornenge
Department of Telecommunication Engineering, Federal University of Technology, Minna, Nigeria

***Abstract*—Community detection in networks is a research area that is gaining a lot of attention mostly because of its great applications in areas such a coding, link prediction, routing, containment of virus and warm online, and recently for criminal detection. In this era of Big Data, it is envisaged that community detection will be handy in solving many societal problems. Many algorithms have been developed to solve the complex problem of clustering groups of nodes in a network. In this paper, a few of these algorithms and their challenges are discussed. Directions for future research in this area are also pointed out.**

Keywords-community detection; social network graphs; binary networks, weighted networks

I. INTRODUCTION

The massive improvement in technology over the years, especially the technical and commercial success of electronic communications devices has made communication and interactions between people very easy. The widely growing attention towards research in big data has led to significant moves and potential moves towards modeling human behaviour and understanding the nature of their relationships through large amounts of data collected over time. A network is a group of nodes (or vertices) connected through edges or links. A community in a network is a group of nodes having more internal connections with each other than external connections with the rest of the network [1]. They are also called Clusters, Cliques or Cohesive groups [2], [3]. Communities can overlap, meaning, members of one community or clique can also be part of another community or clique.

Community detection is commonly carried out with the use of social network graphs. Social Network graphs may be directed or undirected, weighted or unweighted etc. In a directed graph, the direction of communication between two nodes (vertices) is considered, that is, the edges connecting the vertices have directions associated to them. Directed graphs are called digraphs if no multiple edges connect the vertices and multigraphs if multiple edges connect the nodes. Undirected graphs are made up of unordered pairs of vertices, i.e., direction of communication is not important. Also, an undirected graph is unweighted (or binary) if a single edge connects each pair of vertices. In this case, it is only important to know if two vertices communicate or not. The extent or frequency of communication is not important. However, for weighted graphs, there can be multiple edges connecting a pair of vertices, showing that some relationships are stronger than others [4].

In this work, we provide a brief survey on some of the works done in community detection, discuss a few issues and challenges these approaches are yet to completely take care of. We begin by briefly discussing some of the metrics used by researchers for community detection in section II, then we will discuss some of the algorithms and techniques used to detect communities of users within a network in section III. We will do a comparison of these methods in section IV, and conclude by recommending considerations for future research in community detection.

II. METRICS USED IN COMMUNITY DETECTION

A. Centrality Metrics

Centrality measures are used to depict the level of importance or standing of a given node in relation to other nodes in a network or community. These include; Betweeness centrality (BC), which measures the tendency of a node to be found along the shortest path between two other nodes. A node with high BC is important in a network because it serves as an important route for information flow in that network, it means that removal of such node will either collapse the network or weaken it considerable. Closeness centrality (CC) is a measure of the sum of all the shortest paths between that node and other nodes in the graph i.e. it measures how close a node is to other nodes in the network; Degree centrality (DC) measures the level of connectedness of a node i.e., it gives a measure of how many nodes are directly connected to a given node in relation to all other nodes in a network. Eigenvector centrality which shows to what extent a node is connected to other well-connected nodes, this metric gives the intuitive reasoning that an important node will usually be connected to other important nodes. It has a google variant called Pagerank [5][6]. These metrics play important roles in determining which nodes belong to certain communities and how important or influential such nodes are [7].

B. Other Metrics

Other important metrics include Clustering coefficient of a node, which is the probability that any two random neighbours of a given node, chosen at random are, connected; belonging degree, conductance and modularity.

1) Belonging Degree

Assuming C is a community in a network; for a node; $u \in V$; k_u, N_u,are node degrees and neighbor sets respectively. And let w_{uv} be the weight of the link between nodes u and v (where *v* is already in the community).

$$k_u = \sum_{u \in N_u} w_{uv} \quad (1)$$

For the community C, and node u, the belonging degree $B(u, C)$ of node u to community C is defined as:

$$B(u, C) = \frac{\sum_{u \in C} w_{uv}}{k_u} \quad (2)$$

It shows to what extent a given node belongs to a community [4].

2) Conductance

It measures the fraction of total edge volume that point outside the cluster [4]. It measures how well knit a graph is. The lower the conductance value*ϕ(C)*, the more connected the nodes are.

$$\phi(C) = \frac{cut(C, C/G)}{w_c} \quad (3)$$

Where $cut(C, C/G)$represents the number of cut edges in the community (which represents all edges leaving the community), and w_c is the total weight of edges in the community.

3) Modularity

It is based on the idea that a good cluster should have a higher internal and lower external density of edges compared to a *null model* with similar structural properties but without a community structure

$$Q = \sum_{i=1}^{k} (e_{ii} - a_i^2) \quad (4)$$

where, e_{ii} is the fraction of weights of edges belonging to community *i*, while a_i is the fraction of all edges connecting community *i*with other communities.

It is a measure of network partition and it shows the quality of the community structure in the network [8].

III. COMMUNITY DETECTION

The goal of community detection is to partition a network into dense regions of the graph. Each region represents a group of nodes that are closely related, and hence are in the same community. A lot has been done in recent years to detect communities of users from large social or mobile networks. Insecurity issues all over the world have made most of the approaches to find great applications in criminal network detection. While detection could be a little easy in most cases, certain networks, especially covert networks of criminals, contain highly complex structures of communication that cannot be very easily detected. In terrorist networks, for instance, it is possible for one group of people not to know the others [9] e.g. as reported in [10] on the Sept. 9/11 hijackers, those trained to fly may not know the people on ground. This makes the task of detecting and analyzing such networks much trickier using the conventional means of network detection.

Many researchers have undergone the task of coming up with better improved methods to detect communities and analyze behaviour on social graphs. Most of the earlier works were based on binary detection schemes, since many natural networks are biological networks.

A. Binary Networks

Many natural networks are biological networks, which are usually binary. In such networks, it is only important if two nodes are either connected or not connected. The relative strengths of connections between nodes do not matter, a typical binary network is shown in Fig 1.

Most of the earlier algorithms for community detection were based on binary networks. Very prominent among them is one proposed by Girvan & Newman [11] who focused on the boundaries of communities rather than their core. It is said to be the first algorithm in modern age community detection [12]. In their approach, edges are removed from the network based on Betweeness centrality values. The edges with the highest Betweeness centrality are removed, Betweeness is calculated again for the edges affected by this removal, and the process is repeated until no edges remain. Even though it worked considerably well even when tested on real life networks such as the Zachary Karate club [13], it could not be used to detect communities with weighted edges and the run time of the algorithm as the number of nodes increase makes it unsuitable for large graphs.

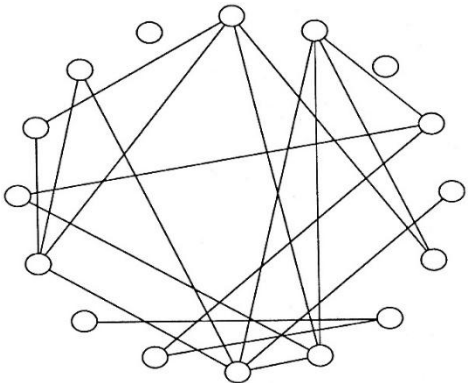

Figure 1. A diagram showing the structure of a binary network

Cfinder was developed to uncover the structure of complex networks by analyzing the statistical features of overlapping networks [14]. In their work, a community (a k-clique community) was defined as a union of all k-cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent k-cliques (where adjacency means sharing nodes). It was based on the fact that members can be reached through well connected subsets of nodes. This approach allowed members of one clique to possibly also be found in another clique (overlapping) which made it better than the divisive and agglomerative methods which form many communities by allowing each node to remain a member of only one community but cuts it off from its other

communities. The community detection was done by setting a threshold weight for the links and ignoring links that were below this threshold weight making it essentially a binary network. The value of the threshold weight is lowered until the size of the largest community is twice as much as the second largest one. Thus making sure that there are no giant communities by merging smaller communities. When the initial community is already unweighted, there is no need for a threshold weight so the smallest value of k for which no giant communities appear is selected.

The RAK algorithm (Name derived from the surnames of the researchers: Raghavan, Albert and Kumara) which is based on label propagation was also proposed in [15]. In their approach each node is first initialized to a unique label which represents the community it belongs to, and these labels then propagate through the network. A node would determine its community based on the labels of its neighbours. Each node joins a community which has the most of its neighbours as members and the labels of the nodes are updated at each iteration. As the propagation continues, dense connected groups of nodes finally settle for a unique label, and in the end, all nodes with the same labels are placed in the same community. This continues until each node in the network has the label to which the maximum number of its neighbours belong to (or one of the labels used by a maximum number of its neighbours), not when the labels no longer change since it is possible to have nodes with equal maximum number of neighbours with two or more labels. Like the Cfinder, this algorithm also worked very well when tested with datasets from the Zachary karate club as well as the U.S. football network. The fact that it is possible for the iteration to end with two disconnected groups of nodes having the same label and requiring a breadth-first search on the subnetwork of each individual group to separate the disjointed communities increases the computation time and complexity of the technique.

Furthermore, in [5] a criminal network graph of 43, 000 nodes was constructed and analyzed from 1,000 publicly leaked email addresses of predominantly Nigerian advanced fee fraud scammers using the Pagerank algorithm. The data was collected by getting the list of friends and friends of friends with the help of some of these email addresses on Facebook. The graph visualization was carried out by a method known as Force Atlas 2 on Gephi. Pagerank algorithm and centrality measures such as Betweeness Centrality, Closeness Centrality, Eigenvector Centrality and Degree Centrality were used to identify key actors in the network and form communities of criminals. Link nodes between communities were also determined through this process. Relative importance of communities could be determined by the number of top players in each community. The method showed 5% possibility of them being scammers and 15% possibility of them being members of scammer communities for a random sample of 100 nodes.

B. Weighted Networks

All of the methods briefly discussed earlier focused on binary networks. In such methods, attributes of nodes are emphasized instead of the edge content which represent the actual link between the nodes. Even though more challenging, edges provide a richer characterization of community behaviour [16]. Most networks are weighted (all connections to a particular node are not equal) so community detection is more reliable when the actual extent of interaction between nodes is considered rather than just friendship between nodes [17].

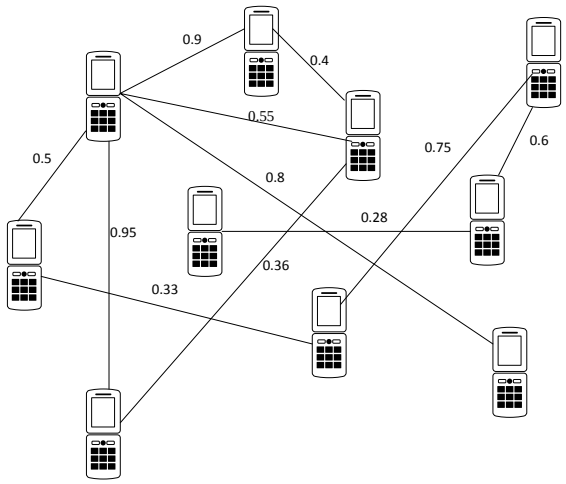

Figure 2. A diagram showing the structure of a weighted network

A notable algorithm for detection of communities in weighted networks is the COPRA algorithm [18]. This algorithm is based on the label propagation algorithm (RAK) discussed earlier. However, in this case, a node can be a member of more than one community, and this algorithm can also handle weighted networks. The method uses a belonging coefficient which shows the strength of a node's membership to a community, which is set to a threshold value $\frac{1}{v}$, where v is a factor in the algorithm which also indicates the maximum number of communities the given node can belong to and the belonging coefficient of each label sum to one. Instead of having just one label, community identifiers are used and a node is allowed to keep more than one community identifier in each label without retaining all of them. During each propagation, the node labels are constructed and the nodes with belonging coefficients less than a given threshold are deleted. If it happens that all the pairs in a node have a belonging coefficient less than the threshold, the one with the maximum belonging coefficient (or one of two or more pairs with maximum belonging coefficient chosen at random) is retained. This algorithm had a huge advantage of allowing overlapping community structures and ability to detect communities on a weighted network. However, just like the RAK algorithm, the COPRA algorithm does not always converge to a constant state where the node labels no longer change after each iteration, thus bringing in more complexity as well.

In [19] a method for finding communities of users by first identifying core nodes and finding cliques around those core nodes was proposed. They argued that having global knowledge of the graph required by most algorithms was unrealistic for very large graphs. A unique feature of their approach is that it is not sensitive to the position of the source node. Their method applied the Girvan & Newman (GN) algorithm [11] to detect all communities in the network, then the nodes with maximal degrees among the communities are found (i.e., core nodes) and the

communities are expanded from these core nodes by finding the most likely nodes closer to the core node using node degrees.

The **Strength** algorithm [8] has a strategy which is to find an initial partial community with maximal node strength and expand by adding tight nodes to the partial community until detection is complete for that particular community. The algorithm consists of two parts: finding initial community, and expanding the community. To find the initial community, the node with the highest node strength (sum of all weights of connections between the node and its neighbours) is chosen along with all its neighbours as an initial community. Any node with a belonging degree less than a threshold (chosen as 0.5) is not connected enough to the community and is thus removed from it. To expand, the belonging degree for all neighbours of the community are calculated. The neighbours with belonging degrees up to the threshold (0.5) are automatically added to the community, while those with belonging degrees less than 0.5 but 0.4 or more are added to the community only if adding them to the community increases the value of the modularity. The expansion stops when no neighbour of the community meets these criteria. It also greatly supports overlapping communities. The algorithm however, degrades in its performance when the overlapping increases.

In [4] a conductance-based algorithm which was an enhancement of the **Strength** algorithm was developed. The algorithm is just like the S**trength** algorithm only that a new objective function, Conductance, is used in addition to the belonging degree, and also, the method of selecting an initial community is slightly different. Here the initial community is a community of two nodes in the network with the highest edge weight between the two of them. To expand, the belonging degrees of all neighbours of the community are calculated and the neighbour with the highest belonging degree (which has the most likelihood of belonging to the community) is temporarily added to the community instead of using a threshold value. If the conductance of this new temporary community is less than that of the initial community, the new node is permanently added to the community and the expansion process is repeated until adding a new node with the highest belonging degree from the neighbours of a community gives a higher conductance than that of the present community. This is based on the fact that a stronger or tighter community has a lower conductance than a weaker one.

IV. SUMMARY

Here, we compare the key algorithms mentioned in the previous sections by looking at their ability to detect communities on weighted graphs, ability to have overlapping nodes in communities, and the technique or metric applied. This summary is presented in Table I.

V. CONCLUSION

This work has briefly discussed community detection in networks. It defined some of the metrics used to measure performances in community detection. It also gave a brief survey of some community detection algorithms in literature, pointing out their strengths and limitations. For example, the result in Table I shows the discussed algorithms used for community detection. Each of these algorithms, as discussed, could be useful depending on the nature of the problem being addressed, and the type or size of network in use as well. However, it was observed that no algorithm so far has detected communities with consideration for the strengths of indirect links between nodes that may either connect directly as well or have no direct connection at all. Research is currently ongoing to explore the possible improvement this may impact on the earlier kinds of networks.

TABLE I. COMMUNITY DETECTION ALGORITHMS

Algorithm	Metric/Technique used	Weighted	Overlapping
Girvan & Newman	Betweeness Centrality	No	No
CFinder	Use of threshold weight	No	Yes
RAK	Label propagation	No	No
COPRA	Label propagation	Yes	Yes
Strength	Belonging Coefficient	Yes	Yes
Conductance-based	Belonging degree and conductance	Yes	Yes

REFERENCES

[1] S. Fortunato, “Community detection in graphs,” *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, Feb. 2010.

[2] S. P. Borgatti, “2-Mode concepts in social network analysis,” *Encyclopedia of complexity and system science*, vol. 6, 2009.

[3] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” Nature, vol. 435, no. 7043, pp. 814–818, Jun. 2005.

[4] Z. Lu, Y. Wen, and G. Cao, “Community detection in weighted networks: Algorithms and applications,” in *Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference on*, 2013, pp. 179–184.

[5] H. Sarvari, E. Abozinadah, A. Mbaziira, and D. Mccoy, “Constructing and Analyzing Criminal Networks,” 2014, pp. 84–91.

[6] M. Rahman, “Application of social networking algorithms in program analysis: understanding execution frequencies,” Colorado State University. Libraries, 2007.

[7] M. Ahsan, T. Singh, and M. Kumari, “Influential node detection in social network during community detection,” in *Cognitive Computing and Information Processing (CCIP), 2015 International Conference on*, 2015, pp. 1–6.

[8] D. Chen, M. Shang, Z. Lv, and Y. Fu, “Detecting overlapping communities of weighted networks via a local algorithm,” *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 19, pp. 4177–4187, Oct. 2010.

[9] S. T. Smith, K. D. Senne, S. Philips, E. K. Kao, and G. Bernstein, “Covert network detection,” *Lincoln Laboratory J*, vol. 20, no. 1, pp. 47–61, 2013.

[10] V. E. Krebs, "Mapping Networks of Terrorist Cells," *Connections*, vol. 24, no. 3, pp. 43-52, 2002

[11] M. Girvan and M. E. Newman, “Community structure in social and biological networks,” *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[12] S. Fortunato and A. Lancichinetti, “Community detection algorithms: a comparative analysis: invited presentation, extended abstract,” in *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, 2009, p. 27.

[13] W. W. Zachary, “An information flow model for conflict and fission in small groups,” *Journal of anthropological research*, pp. 452–473, 1977.

[14] S. Bell, A. McDiarmid, and J. Irvine, “Nodobo: Mobile phone as a software sensor for social network research,” in *Vehicular*

Technology Conference (VTC Spring), 2011 IEEE 73rd, 2011, pp. 1–5.

[15] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, no. 3, Sep. 2007.

[16] G.-J. Qi, C. C. Aggarwal, and T. Huang, "Community detection with edge content in social media networks," in *2012 IEEE 28th International Conference on Data Engineering*, 2012, pp. 534–545.

[17] M. Ovelgönne, A. Geyer-Schulz, and M. Stein, "Randomized greedy modularity optimization for group detection in huge social networks," in *Proceedings of the fourth SNA-KDD Workshop, KDD 2010, July*, 2010, vol. 25, pp. 1–9.

[18] S. Gregory, "Finding overlapping communities in networks by label propagation," *New Journal of Physics*, vol. 12, no. 10, p. 103018, Oct. 2010.

[19] Z. Lu, Y. Wen, and G. Cao, "Community detection in weighted networks: Algorithms and applications," in *Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference on*, 2013, pp. 179–184.