Identification of Subgroups in a Directed Social Network Using Edge Betweenness and Random Walks

K. Sathiyakumari and M. S. Vijaya

Abstract Social networks have obtained masses hobby recently, largely because of the success of online social networking Web sites and media sharing sites. In such networks, rigorous and complex interactions occur among several unique entities, leading to huge information networks with first rate commercial enterprise ability. Network detection is an unmanaged getting to know challenge that determines the community groups based on common place hobbies, career, modules, and their hierarchical agency, the usage of the records encoded in the graph topology. Locating groups from social network is a tough mission because of its topology and overlapping of various communities. In this research, edge betweenness modularity and random walks is used for detecting groups in networks with node attributes. The twitter data of the famous cricket player is used here and network of friends and followers is analyzed using two algorithms based on edge betweenness and random walks. Also the strength of extracted communities is evaluated using on modularity score and the experiment results confirmed that the cricket player's network is dense.

Keywords Edge betweenness · Random walks · Modularity
Community detection · Social network

1 Introduction

The developing use of the Internet has brought about the development of networked interaction environments consisting of social networks. Social networks are graph structures whose nodes represent people, corporations, or other entities, and whose

K. Sathiyakumari (✉) · M. S. Vijaya
PSGR Krishnammal College for Women, Coimbatore 641004, Tamilnadu, India
e-mail: sathiyakumari@psgrkc.ac.in

M. S. Vijaya
e-mail: msvijaya@psgrkc.ac.in

S. C. Satapathy et al. (eds.), *Smart Computing and Informatics*, Smart Innovation, Systems and Technologies 77, https://doi.org/10.1007/978-981-10-5544-7_12

edges represent courting, interaction, collaboration, or have an effect on among entities. The edges in the network connecting the entities may have a direction indicating the flow from one entity to the other, and a strength denoting how much, how often, or how important the relationship is. Researchers are increasingly interested in addressing a wide range of challenges exist in these social network systems.

In recent years, social network studies has been completed the use of massive amount of statistics gathered from online interactions and from explicit courting links in online social community systems such as Facebook, Twitter, Linkedin, Flickr, Instant Messenger. Twitter is highly rated as a new shape of media and utilized in various fields, such as corporate marketing, education, broadcasting. Structural characteristics of such social networks can be explored the usage of sociometrics to recognize the shape of the network, the properties of links, the roles of entities, information flows, evolution of networks, clusters/communities in a network, nodes in a cluster, center node of the cluster/network, and nodes on the periphery, etc. To find out functionally related items from communities, [1, 2] allow us to observe interplay modules, lacking characteristic values and expect unobserved connections among nodes [3]. The nodes have many relationships among themselves in groups to proportion commonplace residences or an attributes. Figuring out network community is a trouble of clustering nodes into small corporations and a node can be belonging to a couple of communities immediately in a network structure.

Unique resources of statistics are used to perform the clustering challenge, first is about nodes and its attributes and the second is ready the relationship among nodes. The attributes of nodes in community structure are known properties of users like network profile, author publication, publication histories which helps to determines similar nodes and community module to which the node belongs. The connection between the nodes provides information about friendships, authors collaborate, followers, and topic interactions.

A few clustering algorithms [4, 5] employ node attributes, however, ignores the relationships among nodes. But the network detection algorithms make use of businesses of nodes which can be densely related [6, 7] but ignore the node attributes. By way of using these sources of records, sure set of rules fails to explain vital structure in a community. For instance, attributes may additionally tell about which community node with few links belonging to and it is far hard to decide from network structure on my own. On the opposite, the community offers detail approximately two nodes belong to identical community even someone of the node has no attribute values. Node attributes can stabilize the community structure which ends up in more correct detection of communities. Thus, community detection becomes difficult undertaking when thinking of both node attributes and network topology.

The proposed method overcomes the above problem by identifying communities based on node and its attributes by implementing Girvan–Newman edge betweenness and random walks algorithm.

2 Related Work

A network is a densely related subset of nodes; this is carefully related to the last network. Social networks are a mixture of essential heterogeneities in complicated networks, together with collaboration networks and interplay networks. Online social networking packages are used to represent and version the social ties among people. Finding communities within an arbitrary community may be a computationally difficult challenge. Numerous research dealings in recent past years have been carried out in the subject matter of network detection, and a number of the crucial research works are point out beneath.

Nicola Barbieri et al. [8] provided network-cascade community (CCN) model, which produced overlapping communities based totally on its interest and level of authority. Major drawback of this model changed into sluggish in getting to know section and also slows in estimate impact energy.

Xie et al. [9] determined numerous lessons of overlapping communities the usage of special algorithms like clique percolation, label propagation [10, 11], agent-based and debris-based totally fashions. This method is used to link partitioning and stochastic generative models.

Evans and Lambiotte [12] used everyday node partitioning to line graph for obtaining hyperlink portioning in authentic community. Ahn et al. [13] used Jaccard coefficient of the neighborhood node to discover similarity among hyperlinks. Kim and Jeong [14] used Infomap approach to encode random stroll path in line network.

Hughes and Palen [15] analyzed how twitter customers react and unfold statistics on social and political issues and located out that those massive activities entice new customers to twitter. Diakopoulos and Shamma [16] accumulated tweets concerning the US presidential election candidates of 2008 and analyzed public emotional response, visible expression, and so on. Kwak et al. [17] analyzed a first rate quantity of twitter dialogues and consumer relationships.

Fortunato [6] affords a entire evaluate in the vicinity of network detection for undirected networks from a statistical physics attitude, while Schaeffer [18] particularly specializes in the graph clustering problem as an unmanaged studying mission. Both surveys in short discuss the case of directed networks; but their consciousness is on the undirected case of the problem.

In this study, the Girvan–Newman algorithm based totally on edge betweenness and random stroll set of rules is applied for coming across communities in networks with node attributes. The twitter facts of the well-known cricket player are taken for take a look at and network of friends and fans is analyzed based on modularity score.

3 Girvan–Newman Algorithm

3.1 Community Detection Framework

The Girvan and Newman is a trendy community finding algorithm. It performs herbal divisions some of the vertices without requiring the researcher to specify the

numbers of groups are present, or placing limitations on their sizes, and without displaying the pathologies obvious within the hierarchical clustering techniques. Girvan and Newman [19] have proposed an set of rules which has three definitive functions: (1) Edges are steadily eliminated from a community, (2) the edges to be removed are selected with the aid of computing betweenness rankings, and (3) the betweenness scores are recomputed for removal of each aspect.

As a degree of traffic float, Girvan and Newman use place betweenness, a generalization to edges of the renowned vertex betweenness of freeman [20, 21]. The betweenness of a part is described because the wide variety of shortest paths between vertex pairs. This amount can be calculated for all edges in the time complexity of o (mn) on a graph with m edges and n vertices [22, 23].

Newman and Girvan [24] define a measure known as modularity, that is a numerical index that shows ideal separation between nodes. For a separation with g corporations, define as $g \times g$ matrix e whose difficulty e_{ij} is the fraction of edges within the authentic network that be part of vertices in group i to those in organization j. Then, the modularity is defined as

$$Q = \sum_i e_{ii} - \sum_{ijk} e_{ij} e_{ki} = \mathrm{Tr}\, e - \left\| e^2 \right\|,$$

which suggests the sum of all factors of x, q is the fraction of all edges that lie inside communities minus the predictable fee of the identical quantity in a graph in which the vertices have the identical degrees; however, edges are positioned at random without look upon the companies. The q = zero indicates that network form is not any stronger than could be anticipated through randomness and values aside from 0 represent deviations from randomness. Limited peaks in the modularity at some stage in the development of the network shape set of rules endorse suitable divisions of the community.

3.2 Girvan–Newman Partitioning Algorithm

Successively Deleting Edges of High Betweenness

Step 1: Find the edge or multiple edges with maximum betweenness; if there can be tie in betweenness, then put off those edges from graph. This system may spilt the graph into numerous additives; it make first degree partition of graph.

Step 2: Recalculate all betweenness values and then remove the edges/edge with high betweenness value. Again split the first-level region into several components such that there are nested within larger regions of graph.

Step 3: Repeat steps (1) and (2) till edges remain in graph.

Computing Betweenness Values

For each node A:

Step 1: Do breadth first search starting at node *A*.
Step 2: Count the number of the range of shortest paths from *A* to every different node.
Step 3: Decide the quantity of waft from *A* to all other nodes.

3.3 Random Walks Algorithm

Random walks is a mathematical idea formalizing a method consisting of a chain of random steps. In case of graphs, given a node that corresponds to a starting point, a random walks is defined due to the fact the collection of nodes fashioned with the resource of a repeating technique beginning from the initial node and randomly transferring to network nodes. At every step, the random walker is positioned on a node of the graph and jumps to a present-day node selected randomly and uniformly among its friends.

Mathematically, allow $GU = (V, E)$ be an undirected graph and v_0 be the starting node of the random walk. At the tth step, the random stroll is located at node i. At $t + 1$ step, the random walks is transferring from node i to node j (neighbor of i) with transition chance $1/k_i$. This defines the transition matrix p of the random walks as

$$\begin{cases} \frac{A_{ij}}{k_i}, & if(i,j) \in E, \\ 0, & otherwise \end{cases} \tag{1}$$

This matrix may be written as $P = D - 1A$, in which $D - 1$ is the inverse of the diagonal degree matrix d. This matrix can also be considered as a diploma normalized model of the adjacency matrix. Random walks are considered to be Markov chains 1, wherein the set of feasible states corresponds to the vertex set of the graph.

Any distribution on a graph G may be represented by the useful resource of a row vector $\pi = [\pi_{1,} \ldots \pi_n]^T$, wherein the ith entry that captures the amount of the distribution is residing at node i. In case of random walks, the chance distribution over the graph g for each node $i \in v$ at any time step offers the opportunity of the random stroll of being at node i. As a result, if π is the preliminary distribution, then $\pi_1 = \pi_P$ is the distribution after one step and $\pi_t = \pi_{Pt}$ is the distribution after t steps. Therefore, it is able to outline a stationary distribution π_s, because the distribution where $\pi_s = \pi_{sPt}$, $\forall t$. The stationary distribution corresponds to a distribution that does not alternate through the years and describes the opportunity that the stroll is being at a selected node after a sufficiently long time. The combination time is the time wanted via the random stroll to reach its stationary distribution. The spectrum of the transition matrix p can be used to sure the combination time of a random walks on a graph and in particular the second largest eigenvalue [25].

4 Experiments and Results

The proposed framework includes four phases: Twitter data, directed network, community detection algorithm, and modularity score. Every phase is described in the following sections, and the architecture of the proposed system is shown in Fig. 1.

A directed network is created using Twitter friends/followers listing as the graph. In this community detection, two algorithms are used for the stage of evaluation of network Girvan–Newman, and random walks algorithm is used to detect communities and subgroups. The size of subgroups is found using Girvan–Newman algorithm and random walks of this network. The algorithm also detects modularity score of community. The real-time data is collected using the twitter application programming interface 1.1 for this research work. Nine thousand records of friends and followers list of the famous cricket player have been crawled from his twitter account. The data is collected at run time from twitter network using R3.2.4, a statistical tool. The cricket player's initial community network as shown in Figs. 2

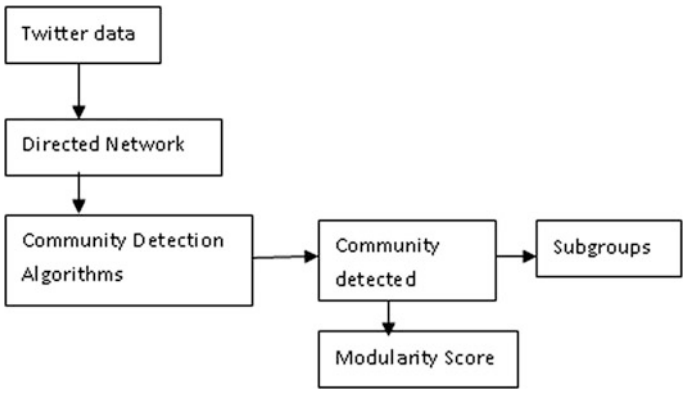
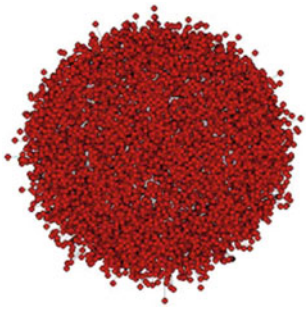


Fig. 1 Community detection framework

Fig. 2 Cricket player's initial network



and 3 depicts the relationship types such as friends, followers, and friends and followers. This network has 7095 edges and 6831 vertices.

A community is a densely related institution of vertices, with only sparser connections to different groups. Girvan–Newman algorithm and random walks algorithm are employed here to detect communities from cricket player's twitter network, since it is a directed network. The modularity score for this network is obtained as 0.91. Thirty-nine different communities are extracted for this network based on edge betweenness modularity measure and demonstrated in different colors as shown in Fig. 4. These 39 communities are clustered based on followers, friends, and both followers and friends in the network. The distribution of nodes in various communities is shown in Fig. 6. The membership of size of community 1 is 69, community 2 has the highest size with 166 memberships. Communities 3 and 4 have the membership sizes 42 and 39, respectively. Communities 5 and 7 have the same membership size 37 and so on. Eight different communities are extracted from the same network based on random walks modularity measure and demonstrated in different colors as shown in Fig. 5. These eight communities are clustered based on followers, friends, and both followers and friends in the network. The distribution of nodes in various communities is shown in Fig. 7. The membership of size of community 1 is 207, community 2 has 166, and community 3 has the highest size with 237 memberships. Communities 3 and 4 have the membership sizes 193 and 209, respectively. Communities 1, 3, and 5 have the high membership size of other

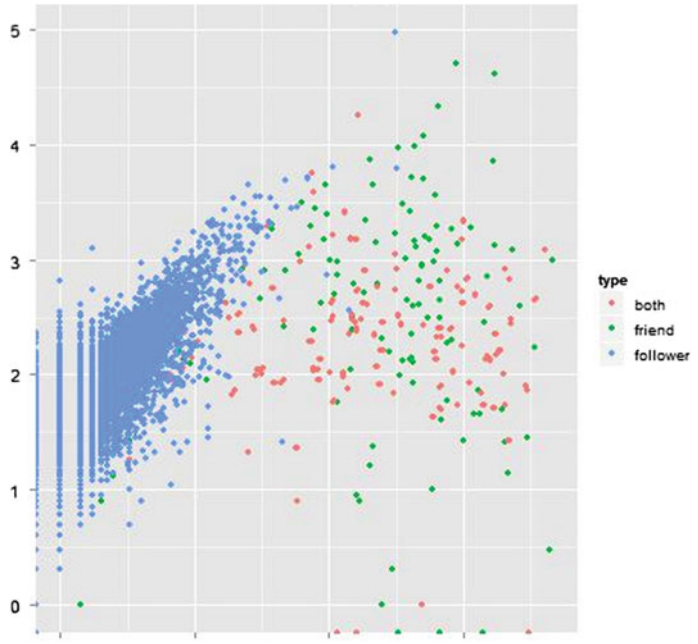


Fig. 3 Friends and followers network

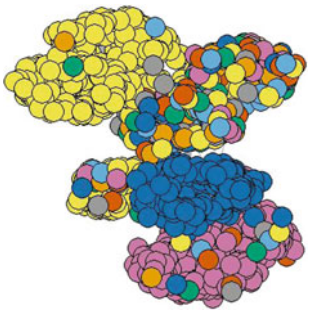


Fig. 4 Communities identified based on edge betweenness algorithm

Fig. 5 Communities identified based on random walks algorithm

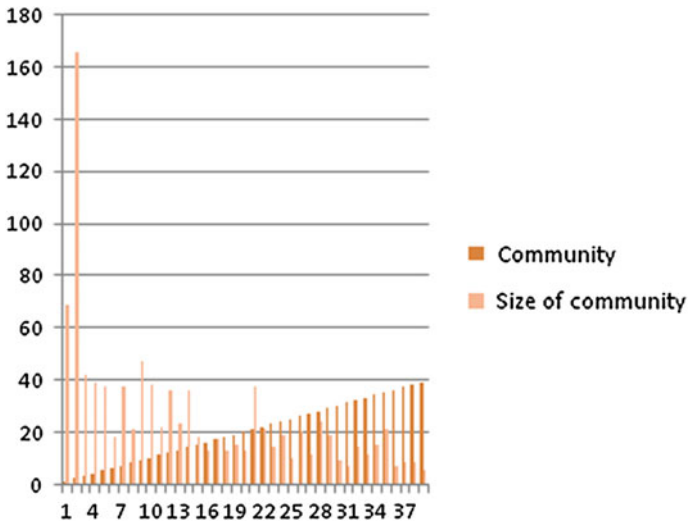


Fig. 6 Membership distribution of communities (edge betweenness algorithm)

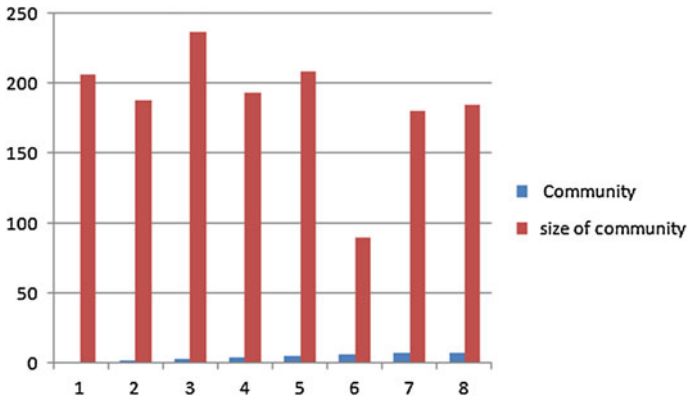


Fig. 7 Membership distribution of communities (random walks community algorithm)

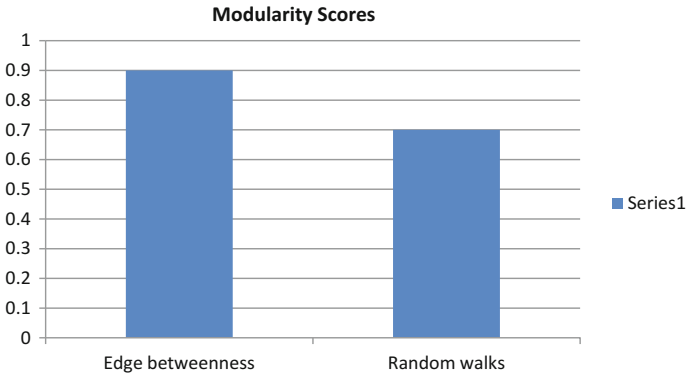


Fig. 8 Modularity scores of community detection algorithms

communities. The modularity score of the network is shown in Fig. 8. The comparative study of community detection algorithms is made in terms of number of communities, membership distribution, and modularity score.

5 Discussion and Findings

The aim of network detection in graphs is to discover the subgroups by using the use of the statistics encoded inside the graph topology. In this research work, the modularity score obtained through edge betweenness algorithm is 0.91, which proves that the cricket player's friends and followers network is dense.

The Girvan–Newman algorithm has detected 39 different communities from the cricket player's network and found five communities dense out of total communities. The modularity score found through random walks algorithm is 0.7, which also confirms that the cricket player's friends and followers network is dense. The random walks algorithm has found eight communities from the cricket player's network which are all highly dense. Girvan–Newman algorithm has discovered more number of sparse communities than random walks algorithm and has eliminated them during clustering. Random walks algorithm finds less number of communities with high communication between the nodes.

6 Conclusion and Future Work

This work elucidates the application of Girvan–Newman algorithm and random walks for detecting communities from networks with node attributes. The real time twitter directed network of a cricket player is used to carry out network analysis. Modularity score is evaluated and subgroups are detected using two community detection algorithms of directed network. The membership distributions of the subgroups generated by two algorithms were discussed. The experimental results indicate that the network is dense and communication between the nodes is high. As scope for further work, analysis of nested communities can be carried out with cliques and subgroups.

References

1. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: Demon: a local-first discovery method for overlapping communities. In: KDD '12 (2012)
2. Girvan, M., Newman, M.: Community structure in social and biological networks. In: PNAS (2002)
3. Yang, J., Leskovec, J.: Overlapping community detection at scale: a non-negative factorization approach. In: WSDM '13 (2013)
4. Blei, D.M., Ng, A., Jordan, M.: Latent Dirichlet allocation. MLR **3**, 993–1022 (2003)
5. Johnson, S.: Hierarchical clustering schemes. Psychometrika **32**(3), 241–254 (1967)
6. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
7. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state of the art and comparative study. ACM Comput. Surv. (2013). doi:10.1145/2501654.2501657
8. Barbieri, N., Bonchi, F., Manco, G.: Cascade-based community detection. In: WSDM'13, February 4–8, Rome, Italy (2012)
9. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state of the art and comparative study. ACM Comput. Surv (2013). doi:10.1145/2501654.2501657
10. Gregory, S.: Finding overlapping communities in networks by label propagation. New J. Phys. **12**(10), 103018 (2010)
11. Padrol-Sureda, A., Perarnau-Llobet, G., Pfeie, J., Munes-Mulero, V.: Overlapping community search for social networks. In: ICDE (2010)

12. Evans, T.S., Lambiotte, R.: Line graphs, link partitions, and overlapping communities. Phys. Rev. E **80**, 016105 (2009)
13. Ahn, Y.-Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
14. Kim, Y., Jeong, H.: The map equation for link communities. Phys. Rev. E **84**, 026110 (2011)
15. Hughes, A.L., Palen, L.: Twitter adoption and use in mass convergence and emergency events. Int. J. Emerg. Manag. **6**(3–4), 248–260 (2009)
16. Diakopoulos, N.A., Shamma, D.A.: Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI '10), pp. 1195–1198 (2010)
17. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or news media? In: Proceedings of the 19th International World Wide Web Conference (WWW '10), pp. 591–600 (2010)
18. Schaeffer, S.E.: Graph clustering. Comput. Sci. Rev. **1**(1), 27–64 (2007)
19. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**, 7821–7826 (2002)
20. Freeman, L.C.: A set of measures of centrality based upon betweenness. Sociometry **40**, 35–41 (1977)
21. Anthonisse, J.M.: The rush in a directed graph. Technical Report BN9/71, Stichting Mathematicsh Centrum, Amsterdam (1971)
22. Newman, M.E.J.: Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. Phys. Rev. E **64**, 016132 (2001)
23. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**, 163–177 (2001)
24. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Preprint cond-mat/0308217 (2003)
25. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)