ISMSIT2017

1st International Symposium on Multidisciplinary Studies and Innovative TechnologiesProceedings Book

(November 2-4, 2017, Tokat, Turkey)

Symposium homepage: www.ismsitconf.org

Community Detection in Social Media Network with Maximum Modularity Using Girvan-Newman Algorithm

Ali Fatih Gündüz[1*] and Ahmet Karadoğan[2]

[1]*Akçadağ Vocational High School, İnönü University, Turkey*

[2]*Computer Engineering Department, İnönü University, Turkey*

**(fatih.gunduz@inonu.edu.tr)*

Abstract –Social networks are formed from interactions of peoples. Measuring the degree of those relationships requires interpreting connectivity of vertices and extracting information from it. Generally individuals form smaller sub-communities in those networks. Identifying those communities by determining sizes of cliques is a challenge and there are numerous solutions in the literature for this problem. In this study we reviewed Girvan-Newman community detection algorithm and applied it on a real life social network obtained from Twitter data. Friendship relations of students of four different universities were used to form the network. A connected graph is generated from this data set in which the students are represented as vertices and followership relations of the students formed the edges of the graph. Since those universities are geographically close to each other, the graph consisted of different link connections among those four clusters. Then community clusters were detected in this connected graph by using Girvan-Newman community detection algorithm.

Keywords –Community detection, Girvan-Newman, data mining, Twitter, social media, social network, clustering

I. Introduction

Social media affects many people from many aspects in our daily life. Message, image, audio, video and other forms sharing occupies a great deal of our time. Starting from short message sharing (SMS) facilities of mobile phones, today social media reached to gigantic dimensions by Facebook, Twitter, Google Plus, LinkedIn and many other online platforms' emergence. For example only Twitter[1] has more than 300 million active users and more than 500 million text messages (a.k.a. tweet) are shared on this platform daily.

Those online platforms paved the way of new social relations. New terminologies like *throwing like, tweeting and being Facebook friends* entered to languages. The activity of social media attract researchers as well. Thanks to them, today we can have thousands of online friends. However Dunbar [1] states that cognitive capabilities of individuals should limit the real size of their network. A normal person is capable having at most 150 friends to interact regularly to sustain a healthy relation. This threshold is known as *Dunbar number*.

Many aspects of social media create interesting study areas such as credibility of messages, community structures, security, natural language processing and etc. Social networks may consist of variable number of cliques. In graph theory, clique term is used to describe a fully-connected subgraph. We usually observe strongly connected partitions even though they are not fully connected. Community detection techniques focus on identifying those partitions.

In this study, we collected Twitter data and examined Girvan-Newman community detection algorithm. Creating a data set from Twitter users who are also students of four Turkish universities, we examined the success of the algorithm on the data set. Although there are many community detection algorithms in the literature, we selected Girvan-Newman algorithm to implement due its intuitive nature.

Community detection algorithms in the literature generally use famous former data sets such as Zachary's Karate Club or Enron email network. Differently, we created a new data set from Gaziosmanpaşa, Sinop, Canik Başarı and Cumhuriyet Universities' students' Twitter accounts. The data set is used to create a graph $G = (V, E)$ where students are represented by the vertices (V) and Twitter followership relations of the students are represented as edges (E).

II. Materials and Method

We used Java to construct data set and then Gephi[2] and Python for graph analysis. Data gathering process, graph construction and Girvan-Newman method will be explained in this section in detail.

A. Data Gathering& Graph Construction

Gaziosmanpaşa, Sinop, Canik Başarı and Cumhuriyet Universities has official Twitter accounts. Those accounts are generally followed by their students for both academic and cultural purposes. Those accounts are respectively followed by 4080, 1492, 1619 and 3381 (totally 10572) Twitter users[3]even though their total number of students are far more

[1]https://analytics.twitter.com/about Accessed October 10, 2017

[2] https://gephi.org/ Accessed October 10, 2017

[3]https://twitter.com/Accessed October 10, 2017

than these numbers[4]. In order to make community detection analysis, we mixed those users together. The universities, their Twitter account names and locations are shown in Table 1.

Table 1. Universities, their official Twitter account and their location

University	Twitter Account	Location
Gaziosmanpaşa	@GaziosmanpasauN	Tokat
Sinop	@sinopuniversity	Sinop
Canik Başarı	@basariuni	Samsun
Cumhuriyet	@cumunivkurumsal	Sivas

We crawled data from Twitter by using Twitter4j API[5]. By querying the twitter account names of the universities we reached to the followers of them which are called users. Then we collect information of each user unless his/her account is blocked. The obtained data consists of unique user id, follower list, description, name, location, favourite count and total number of Twitter statuses. In this phase of the study we queried all followers of each university however blocked accounts are prohibited from accession. Due to the disabilities, we constructed our data set from 1412 different users.

A connected graph is constructed from the data set in which users are represented as vertices and followership relation between users are represented as undirected and unweighted edges. Simply the users following each other are linked by edges. The finally constructed graph consisted of 4424 edges. Average connection per node appeared to be 3.13 which is quite smaller than Dunbar number[1] unfortunately. Having average connection as small as 3.13, the graph consisted of small clusters instead of four large clusters representing the universities.

Moreover since the universities are geographically close to each other, students are likely to be familiar socially. This fact caused intra-cluster links in the graph. By using Gephi tool and algorithm [2], we calculated modularity classes of the nodes. Then visualized it by using Force Atlas algorithm as depicted in Figure 1. As it can be seen in the picture the best modularity score is obtained with cluster number greater than four.

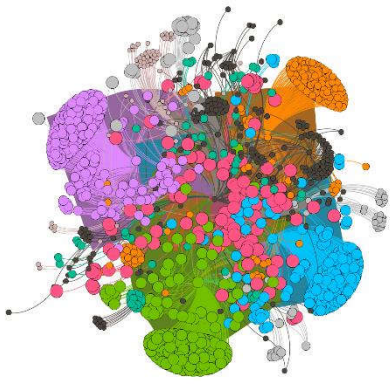

Fig. 1 Gephi based visualization of the graph by using Force Atlas layout algorithm. Clusters are coloured differently.

[4]http://www.yok.gov.tr/web/guest/universitelerimiz Accessed October 10, 2017

[5]http://twitter4j.org/en/index.html Accessed October 10, 2017

B. Girvan-Newman Algorithm

Girvan-Newman algorithm is well known community detection method [3], [4]. The method aims to divide the graph into smaller ones by eliminating the edges between separable clusters. The clustering performance is measured by *modularity (Q)* in this method. Modularity [5], [6] is a metric used to measure division quality of a graph which is formulized below.

$$Q = \sum_i (e_{ii} - a_i^2) \quad (1)$$

$$a_i = \sum_j (e_{ij}) \quad (2)$$

In the Equation 1, *e* is an n by n matrix where n is number of partitions. Its elements e_{ij} are fraction of edges that starts at cluster i and ends at cluster j. On the other hand, in Equation 2, a_i is sum of the columns in the matrix *e*.

Modularity of communities take a value in the range of [0, 1]. A randomly connected graph would have a modularity close to 0. On the other side modularity value of 1 indicates strong inter-community connections. Generally a modularity value in the range of [0.3, 0.7] is expected for natural communities [7].

The separation of clusters is based on *betweenness* values of edges. Betweenness [8] of an edge is defined as the number of shortest paths in the graph which goes through the edge. Girvan-Newman algorithm firstly finds the edges with highest edge betweennes and removes them at each iteration. Algorithm recalculates the edge betweennes to update since the graph structure changes after removal of an edge. Removal of edges arises clique partitions as iterations continue. Like all hierarchical algorithms, community partitions can be described as a dendrogram. The algorithm calculates the modularity values of different partitioning results and finally offer a partitioning version as solution which has the greatest modularity.

To sum up, Girvan-Newman algorithm is simple, intuitive and successful for community detection problem however it has running time disadvantage. When we consider that it requires to find shortest paths between each pair of nodes, its worst running time complexity would be $O(EV^2(V+ E \log V))$ where V is the number of vertices and E is the number of edges. $O(V+E \log V)$ is the complexity of well-known Dijkstra's shortest path algorithm with min-priority queue. All pair combinationscan be calculated with Equation 3.

$$C\binom{V}{2} = \frac{V(V-1)}{2} \quad (3)$$

Since we need to find the shortest paths between each node pairs, there would be $O(V^2)$ shortest path calculations. This procedure would be followed during iterations and in its worst case there would be *E* iterations. Finally when we multiply them, we can estimate Girvan-Newman algorithm's worst case complexity as $O(E* V^2 * (V+ E * \log V))$which is very impractical for big graphs.

C. Obtaining Final Clusters

We aimed to separate the students of the four university from each other. Our preliminary expectation was to obtain

four clusters. However Girvan-Newman algorithm did not give best modularity score for 4 partitions. As it can be seen in Table 2, best total modularity score is obtained with 17 partitions. Depending on the highest probability density value of universities, we combined those 17 partitions into 4 clusters.

$$\text{Max } \{ \text{PDF}(U_i,C_j) = \frac{N_{Ui}}{N_{Total}} \text{ for i=1,2,3 and 4 } \} \quad (4)$$

Pdf value of each university U_i is calculated for each cluster C_j by dividing number of students belonging to U_i (N_{Ui}) by total number of students in the cluster (N_{Total}) in Equation 4. Then clusters are mixed into 4 larger clusters having greatest university pdf.

Purity and entropy values of the mixed clusters are compared with corresponding values of 4 partition Girvan-Newman results in the next section.

III. RELATED WORKS

In 1967 Milgram conducted [9] an experiment to analyse how people are linked to each other. This research put forward the famous "six degrees of separation" theorem and is accepted as the origin of the contemporary social network analysis [10]. Social networks have been studied by many disciplines including economics, sociology, psychology, computer science and etc. There are studies like [10], [11] and [12] that investigate criminal possibilities on social networks and predict counter-terrorism precautions to sustain security.

Observations [13], [14] and [15] of higher clustering coefficients and lower average shortest paths than randomly generated graphs introduced a new phenomenon known as "small world property". Many daily life systems can be seen in the form of networks such as acquaintance, collaboration and friendship networks. Many researcher focused on clustering those networks as well. If a network is capable of being divided into communities then its graph representation can be decomposed into non overlapping vertices sets as well.

There are both agglomerative and divisive hierarchical clustering solutions for this problem [17]. [16] is an example of agglomerative hierarchical clustering algorithm while Girvan-Newman algorithm [3], [4] is a divisive hierarchical clustering technique.

Girvan-Newman method is intuitively plain and simple to comprehend, also it is well-known in the literature [17]. It is based on removing the edges which are most commonly used between vertices. A graph containing numerous communities would inevitably have inter-community edges that are most frequent shortest paths. Iteratively removing them surfaces those communities.

In an another study [18], metaheuristic optimization methods, namely original Bat Algorithm (BA), Gravitational Search Algorithm (GSA), modified Big Bang–Big Crunch algorithm (BB-BC), improved Bat Algorithm based on the Differential Evolutionary algorithm (BADE), effective Hyperheuristic Differential Search Algorithm (HDSA) and Scatter Search algorithm based on the Genetic Algorithm (SSGA) are used to detect communities and their performances are compared.

IV. RESULTS

The best modularity score of Girvan-Newman method is obtained from 17 partitions as it is shown in Table 2.

Purity of clusters are measured according to Equation 5 where K_j is ground truth class j and C_i is cluster i. The purity results are shown in Table 3.Entropy values of the clusters are calculated according to Equation 6. Those results are shown in Table 4. Figure 2 shows the bar chart representation of Table 3 and Figure 3 displays the content of Table 4 in bar chart format as well.

In the tables below, Gaziosmanpaşa, Sinop, Canik Başarı and Cumhuriyet Universities' clusters are indexed as 1, 2, 3 and 4 respectively.

$$\text{Purity} = \frac{TruePositive}{TP+TN+FP+FN} = \frac{1}{n} * \sum_{i=1}^{k} maxj\ |C_i \wedge K_j| \quad (5)$$

$$\text{H}(C_i, \Omega) = -\sum_k P(K_k) \log P(K_k) = -\sum_k \frac{|K_k|}{|C_i|} * \log\left(\frac{|K_k|}{|C_i|}\right) \quad (6)$$

Table 2. Girvan-Newman modularity results for different partitions. Best modularity score line is written bold.

Number of partitions	Total Modularity
2	0.027732
3	0.057407
4	0.075885
5	0.269851
6	0.491687
7	0.505041
8	0.519332
9	0.582215
10	0.634451
11	0.636144
12	0.636580
13	0.637348
14	0.636444
15	0.641735
16	0.642758
17	**0.659686**
18	0.655432
19	0.653654
20	0.653778
21	0.652803
22	0.652771
23	0.650340
24	0.649382
25	0.643628
1412	0.012532

Table 3. Purity scores of partitioning

C_i	Mixed Clusters	Raw Partitioning
1	0.935	0.277
2	0.765	0.100
3	0.538	0.01
4	0.987	0.01

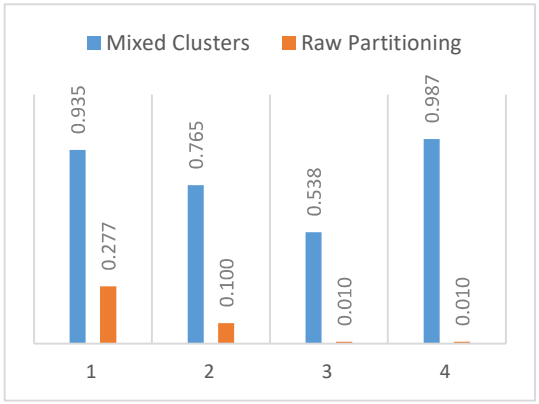

Fig. 2 Purity scores bar chart representation

Table 4. Entropy values of the clusters

C_i	Mixed Clusters	Raw Partitioning
1	0.283	1.277
2	0.546	0.772
3	1.184	3.331
4	0.075	6.661

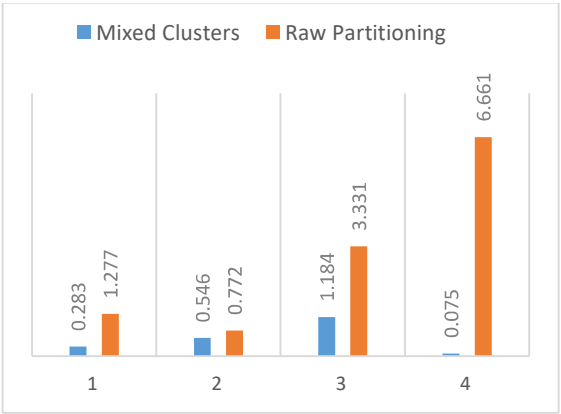

Fig. 3 Entropy scores bar chart representation

V. DISCUSSION

In this section we will interpret the results and compare performance results of mixed clusters with raw partitioning with Girvan-Newman algorithm. Separating the data into 17 partition and then mixing them into 4 clusters instead of using 4 partitioning increased the modularity from 0.075885 to 0.659686. Total modularity increase appeared to be 770%. This uplift is measured in other aspects of clustering performance as well.

As it can be seen in Table 3 and Figure 4, better purity values are obtained for mixed clusters. Final purity of mixed clusters is 0.806 while it is 0.099 for raw partitioning. The increase of the purity is appeared to be 700%. It is also seen that all of the clusters of are purer than raw partitions.

Moreover entropy values of mixed clusters technique are smaller than raw partitioning as it can be seen in Table 4 and Figure 3. Lower entropy observation indicates that those clusters are less dispersed. On the other hand raw partitions have higher entropy indicating their more scattered cluster structure. Average entropy value of mixed clusters is measured as 0.522 whereas raw partitions' average entropy value is calculated as 3.010.

VI. CONCLUSION

To sum up, our study showed that Girvan-Newman algorithm yielded better clustering results as modularity score of partitions increase. Even though our data set is created from four clusters, the internal structure of the graph did not permit its division into four partitions efficiently. Being a real world network representing social relations of the university students, its structure is not exactly predictable.

REFERENCES

[1] Hill, Russell A., and Robin IM Dunbar. "Social network size in humans." Human nature 14.1 (2003): 53-72.
[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000
[3] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Physical Sciences - Applied Mathematics, 2002.
[4] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Phys. Rev. E, 69(2):026113, February 2004.
[5] Newman M. E., Mark E. J., Girvan M., (2004), "Finding and evaluating community structure in networks", Physical Review E, 69 (2), 026113.
[6] M. Newman. 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, vol. 103, no. 23, pp. 85778582.
[7] Gunes, Ismail, and Haluk Bingol. "Community detection in complex networks using agents." arXiv preprint cs/0610129 (2006).
[8] Freeman, L., (1977). A set of measures of centrality based upon betweenness. In Sociometry 40:35-41.
[9] Kleinfeld, Judith. "Could it be a big world after all? The six degrees of separation myth." Society, April 12 (2002): 5-2.
[10] Ressler, Steve. "Social network analysis as an approach to combat terrorism: Past, present, and future research." Homeland Security Affairs 2.2 (2006).
[11] Krebs, Valdis E. "Mapping networks of terrorist cells." Connections24.3 (2002): 43-52.
[12] Xu, Jennifer, and Hsinchun Chen. "Criminal network analysis and visualization." Communications of the ACM 48.6 (2005): 100-107.
[13] Watts, Duncan J. "Networks, dynamics, and the small-world phenomenon." American Journal of sociology 105.2 (1999): 493-527.
[14] de Sola Pool, Ithiel, and Manfred Kochen. "Contacts and influence." Social networks 1.1 (1978): 5-51.
[15] Milgram, Stanley. "Six degrees of separation." Psychology Today 2 (1967): 60-64.
[16] Clauset, A., Newman , M.E.J., Moore, C., (2004). Finding community structure in very large networks. In Physical Review E, 70:061111.
[17] Tasgin, Mursel, Amac Herdagdelen, and Haluk Bingol. "Community detection in complex networks using genetic algorithms." arXiv preprint arXiv:0711.0491 (2007).
[18] Y. Atay, I. Koc, I. Babaoglu, ve H. Kodaz, "Community detection from biological and social networks: A comparative analysis of metaheuristic algorithms", Applied Soft Computing, c. 50, ss. 194–211, Oca. 2017.