Preference relation and Community detection

József Dombi
Institute of Informatics
University of Szeged
Szeged, Hungary
dombi@inf.u-szeged.hu

Sakshi Dhama
Institute of Informatics
University of Szeged
Szeged, Hungary
sakshi@inf.u-szeged.hu

***Abstract*—To study the properties of complex networks we have to generate the artificial networks. Here, we analyze the existing fitness function for the creation of the communities in the synthesis of the artificial network. One of the first benchmarks known as the GN benchmark, it is used for generating networks with non-overlapping communities. In 2008 Lancichinetti–Fortunato proposed a new benchmark (LFR). This benchmark generated networks with overlapping communities. In this paper, we introduce a new preference-based fitness function to assign a membership relation to the nodes. Here, we introduce a novel approach for controlling the creation of overlapping structures by using preferences. The proposed method works well if there are less than ten percent outliers when tested on smaller graphs of size ≈ 15 to ≈ 500**

***Index Terms*—Preference relation** Overlapping Structure community detection outliers complex networks

I. Introduction

All the things around us can be connected using a structure of networks. Today social networks are at the center of research. The history of social networks began in early twentieth century.The World Wide Web has connected people all over world[14]. Earlier there was point-to-point communication, which was quite limited. The use of social media in the last decade has created new challenges such as data storage, browsing speed, security, privacy, fake news, advertisement and announcing events etc. These new and ever changing challenges require novel solutions[7].

The structural properties of complex networks are of great interest to researchers. The real-world social networks may be represented by undirected or directed graph with positive weights on its edges where the number of vertices is much larger than the edges. These social networks have small world network properties with a large clustering coefficient[4]. The average path length between two nodes is of the order $\ln N$ in a random graph. The network consists of nodes with meaningful relationships among them in the form of links[3]. At another level the relation hierarchy can defined by another way, where the elements of hierarchy are communities.

The structure of these communities is one of the most important properties on these networks. The communities denotes the group of objects that interact with each other in the network.

Intuitively, the definition of the community is a sub-graph where the number of edges inside the sub-graph is more than the number of edges outside the sub-graph[9]. We call the community strong when each node has more connections with the nodes inside the community than outside the community. So a community consists of more connections inside than outside. As communities consists of nodes with more connections within the community.

Real-world networks are different from artificial networks. In real-world networks, the communities have more connections due to relations in the structure of the network. In such networks the power law implies that some hubs or vertices have many more connections than rest of the vertices in the network[9][6] Figure 1 shows the sociometry of real graph commonly known as the Moreno network[7]. This network consists of 33 nodes and it was created by studying the behaviour of grade four students in a school. This network was created to study the friendship pattern among the students. One of the deciding factors was the gender. It used triangles to denote the boys and circles to denote the girls. Using the method of sociometry it visualized these connections in several networks called sociograms [1]. The network could use gender to determine the relation of friendship. However, besides these groups it is possible that there is a community structure within the same gender. On a broader note we use other characteristics of the networks to detect the communities. For example, using the walk-trap algorithm [10] we found 6 communities in the network (see fig 2). To study this concept and its properties for large networks($\approx 10^6$ nodes or more), we present a new method, that defines communities on artificial networks. It is closer to real-world communities on graphs. In figure 2, we calculate 6 communities on the same network using walk-trap algorithm.

II. Related work

Over the time, a number of methods have been developed. One of them is the minimum-cut method for community identification. With this method, the network is partitioned in such a way that it minimizes the edges between the partition. The disadvantage of this method is that communities have a fixed size and they are identified in the network by this predetermined number.

The project was supported by the European Union and co-funded by the European Social Fund (no. EFOP-3.6.3-VEKOP-16-2017-0002). This research was partially supported by grant TUDFO/47138-1/2019-ITM of the Ministry for Innovation and Technology, Hungary.

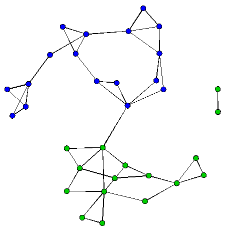

Figure 1. Moreno network Sociometric

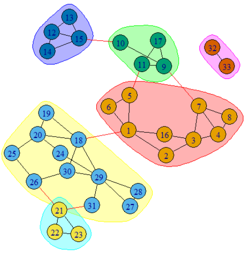

Figure 2. Moreno network walk-trap communities

Another algorithm is called GN(Girvan–Newman) algorithm.The GN algorithm uses edges which are highly localized in the communities [5]. The algorithm identifies these edges present between communities. Some of the edges are rewired and some are deleted. The GN algorithm has a simple structure, but it does not identify the overlapping structures in the network[8]. Therefore for those graphs where overlapping structures are present, LFR (Lancichinetti–Fortunato–Radicchi) benchmark is used.

The LFR benchmark [12] for apriori known communities generates the network which have nodes and communities with a power law distribution. As stated in the "detecting overlapping structure in hierarchical networks ",a bigger value of overlapping structures defines a fuzzier community structure. As stated by the LFR benchmark, the communities are identified as the maximization of a property or fitness of node.The fitness function for a subgraph is defined as

$$f_{\mathcal{G}} = \frac{K^{\mathcal{G}}_{in}}{(K^{\mathcal{G}}_{in} + K^{\mathcal{G}}_{out})^{\alpha}} \tag{1}$$

where $\mathcal{G}$ is the module,$K^{\mathcal{G}}_{in}$ is total internal degree (i.e. double the number of links), $K^{\mathcal{G}}_{out}$ is the total number of links of each member to the rest of the graph and α is a positive real-valued parameter which controls the size of the communities.

The authors of this algorithm [13] intended that the subgraph starting from a node when combined with a new node would have a higher fitness function value and the removal of any node from the sub-graph would lower the value of the fitness function. To compare the partition generated by this fitness function, they used normalized mutual information.

A. *Preference Relations*

The Relation R between X and Y in Set theory is defined as the subset of Cartesian product (i.e. $R \subset X \times Y$), where $X \times Y$ is set of all ordered pairs (x, y) and $x \in X, y \in Y$.

To define the preference, we take a simple example. Among a set of items X if we make or express out preferences by making comparisons of form, ”I strictly prefer x to y”. Preference or choosing x is different than liking x or having taste for x. One can prefer x to y but dislike both the options. The preference relation tell us how true $(x < y)$ is. That is,

$$P^{\nu}_{\alpha}(x, y) = degree‘(x < y) \tag{2}$$

which may be true if
$P(5 < 7) = 0.9$
$P(5 < 5) = 0.5$
$P(7 < 5) = 0.1$
$P(x, y) \in (0, 1)$ in Boolean algebra $P(0, 1) = 1$ and $P(1, 0) = 0$.

We can also define our own function. Here, the intensity of the preference is controlled by the parameter of this function, and α
is the sharpness parameter.

The parameter $\nu \in (0, 1)$ and f is a generator of a strict t norm.The preference implication in pliant logic form is

$$P^{(\alpha)}_{\nu} = \begin{cases} 1, & \text{if } (x, y) \in (0, 0), (1, 1) \\ f^{-1}\left(f(\nu)\frac{f(y)}{f(x)}\right), & \text{otherwise} \end{cases} \tag{3}$$

III. Preference relation for Communities

In order to realize the goals set forth we have constructed the fitness function equivalent of a Preference relation function for creating an artificial network like this[2]. Our system allows us to control the strength of truth when including a new member to the group or community[15]. A node can have the characteristics of more than one community and in such a case it is a harder to decide the membership of the node. The intensity of the preference implication[11] offers a solution to this problem.

We start with a given graph and an initial number of communities required to define the communities on this graph. While comparing the fitness value of the community when a new member is approaching the community to gain its membership, it is important to check how beneficial this would be to the community. This decision is much harder to make when many nodes wish to be member of the community but the community restrics the number of new people that can be members. That is,

$$\mathcal{SG}' = \mathcal{SG} \cup A \tag{4}$$

where $\mathcal{SG}$ is the community graph and A is new member.$\mathcal{SG}'$ is the fitness of the community graph with A. $\mathcal{SG}$ is the fitness of the community graph without addition of a new member A. The selection of a suitable threshold value for a community membership remains open because it cannot be decided merely by calculating the difference between the two. It remains a multi-criteria decision making problem. Each community has some links within the community that includes the connection only among the members of the communities, which is k_{in}, and k_{out} are the links with nodes outside the community.

```
Initialize $community \leftarrow$ initial node;
Initialize threshold;
foreach $c \in community$ do
    Find its neighboring nodes
    foreach $i \in neighborhoodofc$ do
        Calculate the fitness value and preference value
        $X : fitnessofG,$
        Y : fitness of G', where G' = G + i,
        $PreferenceListOfc_i : P_\nu^{(\alpha)}(x, y) = \frac{1}{1 + \frac{1-\nu}{\nu}\left(\frac{1-y}{y}\frac{x}{1-x}\right)^\alpha}$
    end
    foreach n in PreferenceListOfc do
        if $PreferenceList0fc_i > \delta$ then
            $i^{th}$node is member of the community c
        end
    end
end
```

A. Algorithm

The algorithm follows the following steps.

IV. Our Approach

The stopping criteria of the algorithm is decided by this threshold until there is a change in the community size. The members in the neighborhood are approached for membership assignment. We define the initial node of the community as the starting node for each community for the process of community detection. These starting nodes are different from cluster heads as they serve only as the starting criteria for the algorithm and remain fixed for the entire process. As regards the LFR benchmark [12], each node in the neighborhood of these initial nodes is considered for membership based on the fitness value of community. However, because the community is intially small in size (i.e. having only one initial node), the nodes in the neighborhood have a low threshold to join the community. From this viewpoint, the user can control the threshold ν for each community at different stages of the membership allocation. The membership of a node is highly dependent on the membership of the neighboring nodes. The parameter ν has different values when examined on the plot.

V. Results

For a simulation we took six networks to test the feasibility of our approach. The first test consisted of six artificial networks with a similar degree of distribution.

The networks on which our proposed method were tested were generated from the LFR benchmark [12]. For simplicity we took network with nodes with a maximum membership of two communities. II describes network with different sizes and similar statistics. On these networks, we used our proposed method to detect the community membership for nodes. We initially chose the number of communities between 0.1 and 0.2 on smaller graphs and a smaller value from 0.05 and 0.1 on larger graphs. However, the initial number of possible communities depends on the type of network being studied. As can be seen, the algorithm is executed for a fixed number of steps. We observed that on smaller graph need a bigger number of initial communities to reduce the number of outliers in output results. Below, 3 shows on different types of networks investigated in our experiment.

Table I
Artificial network statistics from LFR benchmark where the mu mixing parameter is 0.2 for each network. The membership of overlapping nodes is restricted to 2 .Here N: Number of nodes, NC:size of smallest community,MC2: size of largest community,k: average degree,k_{max}: maximum degree,ON: Number of overlapping nodes,E: Number of edges,T_{trans}: Clustering coefficient,D: Diameter

N	k	k_{max}	NC	MC	ON	E	T_{trans}	D
15	3	5	3	5	5	37	0.3214286	5
30	6	10	5	10	10	158	0.4263646	4
60	12	20	5	10	10	709	0.4819797	3
100	5	15	10	30	15	565	0.2453694	4
200	20	40	5	40	40	3900	0.4070859	3
500	20	50	5	50	50	9962	0.4444339	4

Table II
Preference based relation method results.Here N: Number of nodes, NC:size of smallest community,MC:size of largest community,comNum: Number of communities created in the network,Outliers- Nodes which do not belong to any community

N	NC	MC	ComNum	ON	outliers
15	4	7	5	8	2
30	6	12	6	20	4
60	13	22	10	52	1
100	7	14	20	57	10
200	21	44	16	161	17
500	55	19	32	303	33

VI. Conclusion and Future Work

Here, we presented a new approach for the selection of community membership based on preference. The preference-based approach assigned a membership to the nodes in the

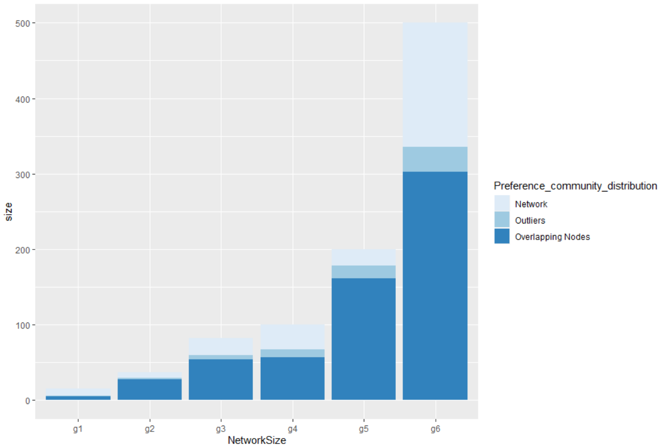

Figure 3. Bar plot shows the number of overlapping nodes and outliers in different sized network

network. Our findings revealed that the distribution of nodes in different communities is not possible without outliers. To control the outliers we did not make any changes to the overall network by adding or deleting any edges. For the networks which are not robust, this way of community detection can be useful as it maintains the original structure of the graph at the end of the algorithm. The distribution of community membership can be controlled by the δ parameter in our proposed method. The preference-based approach provides a new direction for analyzing the overlapping region of communities in networks. The parameter ν controls the threshold and controls the sharpness of the preference. In the future, we would like to perform these tests on a wider range of networks with different structures to gain a better insight into the network properties and the preference-based relation for communities.

References

[1] Jacob Levy Moreno. "Who shall survive?: A new approach to the problem of human interrelations." In: (1934).

[2] József Dombi. "Basic concepts for a theory of evaluation: the aggregative operator". In: *European Journal of Operational Research* 10.3 (1982), pp. 282–293.

[3] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994.

[4] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: *nature* 393.6684 (1998), p. 440.

[5] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.

[6] Roger Guimera et al. "Self-similar community structure in a network of human interactions". In: *Physical review E* 68.6 (2003), p. 065103.

[7] Mark EJ Newman. "The structure and function of complex networks". In: *SIAM review* 45.2 (2003), pp. 167–256.

[8] Leon Danon et al. "Comparing community structure identification". In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.09 (2005), P09008.

[9] Gergely Palla et al. "Uncovering the overlapping community structure of complex networks in nature and society". In: *nature* 435.7043 (2005), p. 814.

[10] Pascal Pons and Matthieu Latapy. "Computing communities in large networks using random walks". In: *International symposium on computer and information sciences*. Springer. 2005, pp. 284–293.

[11] J Dombi, Zs Gera, and N Vincze. "On Preferences Related to Aggregative Operators and Their Transitivity". In: *LINZ* (2006), p. 56.

[12] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. "Benchmark graphs for testing community detection algorithms". In: *Physical review E* 78.4 (2008), p. 046110.

[13] Andrea Lancichinetti and Santo Fortunato. "Community detection algorithms: a comparative analysis". In: *Physical review E* 80.5 (2009), p. 056117.

[14] Linton C Freeman. "The development of social network analysis–with an emphasis on recent events". In: *The SAGE handbook of social network analysis* 21.3 (2011), pp. 26–39.

[15] József Dombi and Tamás Jónás. "Approximations to the Normal Probability Distribution Function using Operators of Continuous-valued Logic". In: *Acta Cybernetica* 23.3 (2018), pp. 829–852.