Horizon | Journal of Humanity and Artificial Intelligence

DIVISION OF SOCIAL NETWORKS INTO TWO COMMUNITIES USING THE MAXIMUM LIKELIHOOD METHOD

Dilshodbek Zakhidov
Senior Lecturer of TMC Institute
mail: birinchi_dilshod@mail.ru

Usarov Jurabek
Senior Lecturer of TMC Institute

Abstract

Identification of communities in graphs is the task of dividing graph nodes into groups (communities) based on the structure and connections between them. This is an important task in social network analysis, bioinformatics, physics and graph theory. There are many methods for extracting communities in graphs. Each method has its advantages and disadvantages and can be used depending on the specific task.

The development of new methods for extracting communities in graphs is an active area of research. For example, the maximum connectivity method and the tree-based clustering method have been developed recently and have shown high accuracy of community extraction on social network and bioinformatics data, respectively.

Comparative analysis of various methods for identifying communities in graphs can help you choose the most appropriate method for a particular problem. This analysis may include factors such as the accuracy of community extraction, runtime, and the ability to work with large datasets.

When social networks are divided into groups, it is important to find the most realistic situation in them. If a division into 2 groups is supposed, that is, if a group is divided into 2 in a social network, then the problem of predicting how these groups will look like is considered in the case of a dodecagonal network. The division into teams was calculated using the Maple program.

Keywords: maximum likelihood. Graphics. Communication between teams. Teams section. Maple.

INTRODUCTION

Social networks play an important role in modern society, connecting millions of people around the world. However, in some cases, these networks can be divided into two groups according to the maximum likelihood method.

The maximum likelihood method is a statistical approach that is used to find the optimal split of some sample. In the context of social networks, this method can be used to divide users into two groups with opposing opinions.

The first step in applying the maximum likelihood method to social media data is to estimate the likelihood of a connection between two users. In social media, bonding is usually determined by mutual friends or each other's followers, although there may be other criteria such as interactions or shared interests.

Then, using these connection probabilities, an adjacency matrix can be constructed that describes the degree of interaction between all users in the network. This matrix can then be used to split users into two groups to maximize likelihood.

The result of this approach is the division of users into two groups, which can be called "conservative" and "liberal". The conservative group are those users who tend to maintain the status quo and preserve traditional values. The liberal group are users inclined to develop and adopt new ideas and approaches.

It is important to note that this approach is not universal and can only be applied to certain types of social networks where it is possible to measure the likelihood of interaction between users. In addition, grouping results can be controversial, and the choice of partitioning threshold can greatly influence the partitioning results.

In general, dividing social networks into two groups using the maximum likelihood method is an interesting approach to understanding the mentality of different users. It can help us better understand how groups of users interact with each other on a social network, and how these interactions can influence social dynamics in general.

A probabilistic approach, called the maximum likelihood method, widely used in mathematical statistics, can be used to identify communities in a network. Following the approach described in [1], we will write a mathematical model for community detection based on the maximum likelihood method.

It is clear that the tightness of relations within society is higher than outside society. We consider the following parameters: The probability of a connection between any two vertices within a team is p_{in}, the probability of a connection between two vertices from different teams is p_{out}.

Consider a network $G=(X,Y)$ in which the set of vertices has the form $X=\{1,2,\dots,n\}$. The number of edges of the network is $m=m(Y)$. Let the connection between vertices i and j be as follows:

$$E(i,j)=\begin{cases}1, \text{If there are connections between vertices i and j,}\\ 0, \text{If there is no connection between vertices i and j.}\end{cases}$$

By a community S we mean a non-empty subset of network vertices, and by a partition $\Pi(X)$ we mean a set of non-overlapping communities whose union is exactly the set X: $N:\Pi(X)=\{S_1,S_2,\dots,S_k\}$, where $\bigcup_{k=1}^{K} S_k, k=1,...,K$.

Assume that the real partition of the network is $\Pi(X)=\{S_1,S_2,\dots,S_k\}$. Let the variables $n_k=n(S_k)$ and $m_k=m(S_k)$ denote the number of vertices and edges in the community $S_k, k=1,...,K$, respectively. Then $n=\sum_{k=1}^{K} n_k$ and $\sum_{k=1}^{K} m_k \le m$.

Let us express the conditions under which the division into teams is optimal.

Let's look at the community $S_k \in \Pi$. The probability of creating m_k connections between n_k vertices in the S_k community is p_{in}.

Each vertex i in the community S_k can have $n-n_k$ connections to the vertices of other communities, but in fact it is connected to the vertices of other communities $\sum_{j\notin S_k} E(i,j)$ has connections.

The probability of realizing a network with a given structure is

$$L_\Pi = \prod_{k=1}^{K} p_{in}^{m_k}\left(1-p_{in}\right)^{\frac{n_k(n_k-1)}{2}-m_k} \prod_{i\in S_k} p_{out}^{\frac{1}{2}\sum_{j\notin S_k} E(i,j)} \left(1-p_{out}\right)^{\frac{1}{2}\left(n-n_k-\sum_{j\notin S_k} E(i,j)\right)} \qquad (1)$$

Taking the logarithm of the likelihood function L_Π (1) and simplifying it, we get

$$l_\Pi = logL_\Pi = \sum_{k=1}^{K} m_k logp_{in} + \sum_{k=1}^{K}\left(\frac{n_k\left(n_k-1\right)}{2}-m_k\right)log\left(1-p_{in}\right)+$$

$$+\left(m-\sum_{k=1}^{K} m_k\right)logp_{out}+\left(\frac{1}{2}\sum_{k=1}^{K} n_k\left(n-n_k\right)-\left(m-\sum_{k=1}^{K} m_k\right)\right)log\left(1-p_{out}\right) \qquad (2)$$

The partition Π^* for which the function l_Π reaches its maximum over all possible partitions is called *optimal*. Note that there is still uncertainty in the choice of probabilities p_{in} and p_{out}. The function $l_\Pi = l_\Pi\left(p_{in}, p_{out}\right)$ depends on the arguments p_{in}, p_{out}. Maximizing l_Π with respect to p_{in}, p_{out} , one can then use these values in numerical calculations.

Statement. For a fixed partition Π, the function $l_\Pi = l_\Pi\left(p_{in}, p_{out}\right)$ reaches its maximum at

$$p_{in} = \frac{2\sum_{k=1}^{K} m_k}{\sum_{k=1}^{K} n_k^2 - n}, \quad p_{out} = \frac{2\left(m-\sum_{k=1}^{K} m_k\right)}{n^2-\sum_{k=1}^{K} n_k^2}. \qquad (3)$$

Substituting (3) into (2), we obtain an expression that depends only on the structure of the network. The given values of the parameters also maximize the likelihood function (1).

If we divide social networks into two groups, we will study which one is more realistic.

II. Main part.

Numerical experiments. Let's look at different types of hexagon social network. We divide the given network into 2 groups and find the one that is most similar to the truth. When the hexagon divides the network into 2 teams, the following situations occur:

№	1st team	2nd team
1	2 edges	4 edges
2	3 edges	3 edges

First, let's look at the following graph:

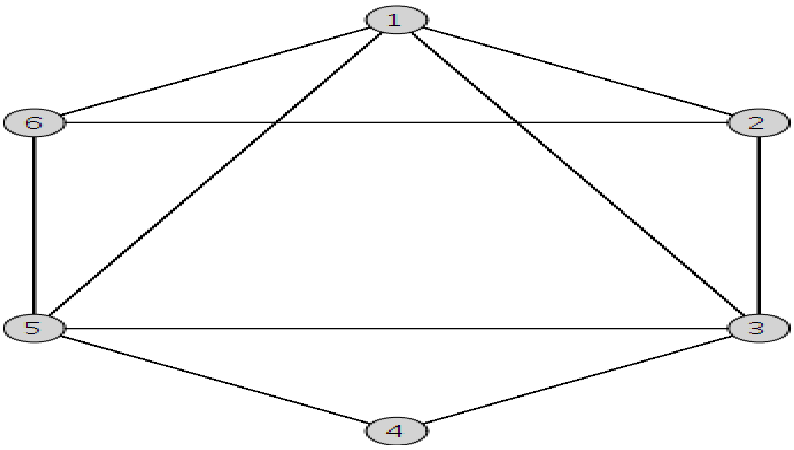

Fig.1 is a network with 6 vertices (10 edges).

This network has 12 vertices and 32 edges. Let's calculate the value l_Π for division in the first case.

We obtain the probability function (2) for the partition

$$\Pi = \{1,2\} \cup \{3,4,5,6\}$$

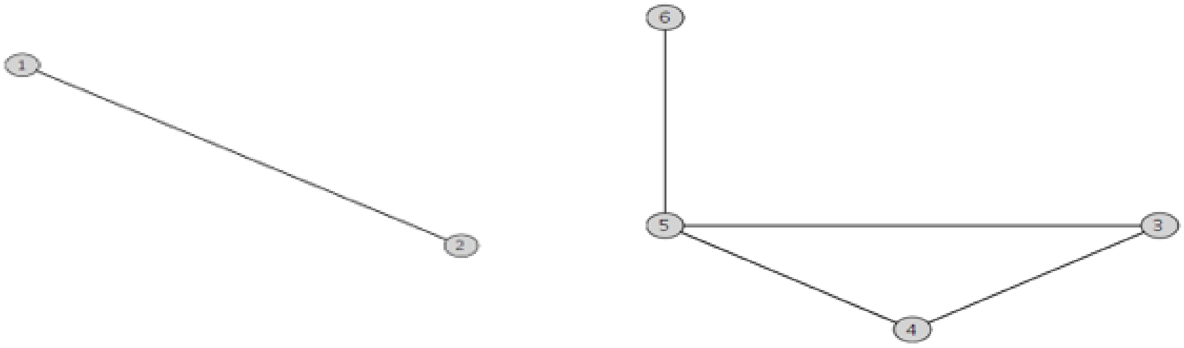

Fig. 2 division of the network into two parts

The total number of vertices is $n=6$, and the total number of edges is $m=10$, since there are 2 teams, the first team has 2 vertices and 1 edge, and the second team has 4 vertices and 5 edges.

$$l_\Pi = 5logp_{in} + 2log(1-p_{in}) + 5logp_{out} + 3log(1-p_{out})$$

We differentiate the function l_Π with respect to p_{in} and p_{out} and set the derivative to 0.

$$\begin{cases} \frac{5}{p_{in}} - \frac{2}{1-p_{in}} = 0 \\ \frac{5}{p_{out}} - \frac{3}{1-p_{out}} = 0 \end{cases}$$

Its maximum is reached at $p_{in} = \frac{5}{7}$ and $p_{out} = \frac{5}{8}$ and is equal to -9.480393026 .

Let's calculate the value of l_Π for dividing $\Pi = \{1,2,3\} \cup \{4,5,6\}$ in the second case.

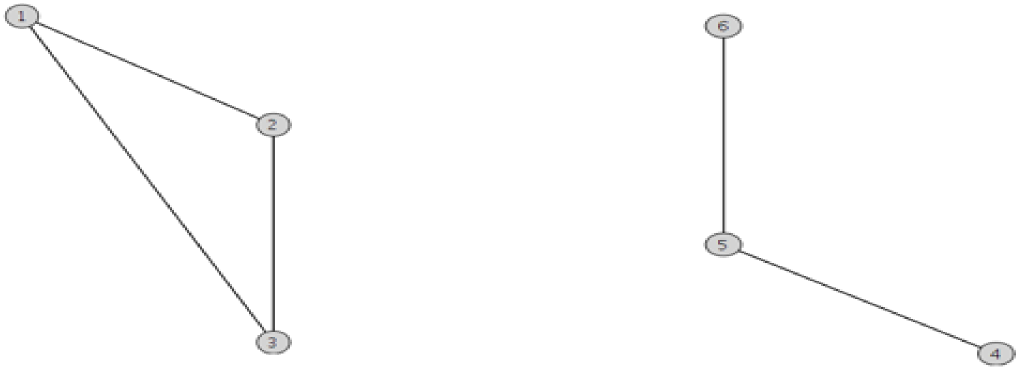

Fig.3. dividing the network into two parts

The total number of vertices is $n=6$, and the total number of edges is $m=10$, since there are 2 teams, the first team has 3 vertices and 3 edge, and the second team has 3 vertices and 2 edges.

$$l_{\Pi}=5logp_{in}+log\left(1-p_{in}\right)+5logp_{out}+4log\left(1-p_{out}\right)$$

We differentiate the function l_{Π} with respect to p_{in} and p_{out} and set the derivative to 0.

$$\begin{cases}\frac{5}{p_{in}}-\frac{1}{1-p_{in}}=0\\ \frac{5}{p_{out}}-\frac{4}{1-p_{out}}=0\end{cases}$$

Its maximum is reached at $p_{in}=\frac{5}{6}$ and $p_{out}=\frac{5}{9}$ and is equal to -8.886021442.

Now imagine all the divisions in the first case:

№	Разделение	l_{Π}	p_{in}	p_{out}
1	$\Pi=\{1,2\}\cup\{3,4,5,6\}$	-9.480393026	$\frac{5}{7}$	$\frac{5}{8}$
2	$\Pi=\{2,3\}\cup\{4,5,6,1\}$	-9.480393026	$\frac{5}{7}$	$\frac{5}{8}$
3	$\Pi=\{3,4\}\cup\{5,6,1,2\}$	-8.415991673*	$\frac{6}{7}$	$\frac{1}{2}$
4	$\Pi=\{4,5\}\cup\{6,1,2,3\}$	-8.415991673*	$\frac{6}{7}$	$\frac{1}{2}$
5	$\Pi=\{5,6\}\cup\{1,2,3,4\}$	-9.480393026	$\frac{5}{7}$	$\frac{5}{8}$
6	$\Pi=\{6,1\}\cup\{2,3,4,5\}$	-9.480393026	$\frac{5}{7}$	$\frac{5}{8}$
max		-8.415991673		

This table shows that the maximum value of l_{Π} is -8.415991673.

When l_{Π} reaches its maximum value, its p_{in} is large compared to others and its p_{out} is small compared to others.

When dividing 3 by 3, i.e. when the 1st team has 3 vertices and the 2nd team has 3 vertices, the maximum likelihood function, l_{Π} , reaches its maximum when dividing $\Pi = \{3,4,5\} \cup \{6,1,2\}$, which has a value of -6.182654189.

we can do these calculations using the algorithm in the "maple" program presented in [2].

CONCLUSION. If we compare the maximum value of l_{Π} in each state, we see that $l_{\Pi} = -6.182654189$ in division of 3 by 3 is greater than the rest, the p_{in} it gets is greater than the p_{in} in the other state, and p_{out} we see that it is small compared to other cases.

To conclude, if the above network with 6 vertices and 10 edges is divided by two, its likelihood function l_{Π} reaches its maximum value when dividing 3 by 3, that is, in the following case:

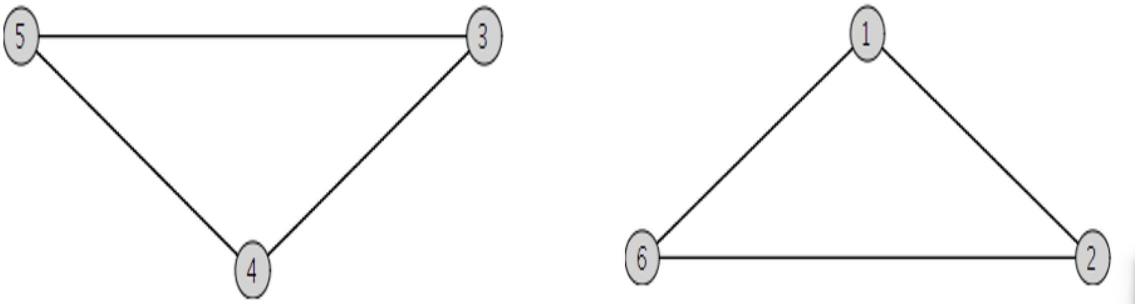

Fig.4. dividing the network into two parts

References

1. Мазалов В. В., Никитина Н. Н. Метод максимального правдоподобия для выделения сообществ в коммуникационных сетях // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2018. Т. 14. Вып. 3. С. 200–214. https://doi.org/10.21638/11702/spbu10.2018.302.
2. Dilshodbek, Z., & Bektosh, S. (2023). THE MAXIMUM REALIZATION METHOD OF COMMUNITY GROUPING IN SOCIAL NETWORKS. CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES, 4(5), 56-61. https://doi.org/10.17605/OSF.IO/5RQ2S
3. Fortunato S., Barthelemy M. Resolution limit in community detection. Proceedings of the National Academy of Sciences USA, 2007, vol. 104(1), pp. 36–41.
4. Girvan M., Newman M.E. J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences USA, 2002, vol. 99(12), pp. 7821–7826.
5. Copic J., Jackson M., Kirman A. Identifying community structures from network data via maximum likelihood methods // The B. E. Journal of Theoretical Economics. 2009. Vol. 9, iss. 1. P. 1635–1704.