2015 APMCM PROBLEMS

Problem C

Identifying the error connections in the network

The network is a powerful tool to describe the structure of a real system—— the social network describes the relationship between human beings, and the World Wide Web describes the hyperlink relationship between web pages. With the development of modern technology, we have accumulated more and more network data, but the data is partially incomplete, inaccurate or sometimes distorted. For example, in the biological network, some early proved existing gene-gene and protein-protein interrelations are overturned by new experiments with higher accuracy.

This topic will address real network problems from biology, information and social networks with data of 6 networks. The scale of these networks is ranging from hundreds of nodes to millions of nodes. Each network connection may be undirected (for example, friend-connection in twitter), or directed (such as people "follow" others in twitter). Based on the original real network, we have added a number of false connections which meet following criteria: (1) the number of the false connections is not more than 10% of the total number of connections; (2) the error connections are picked in a completely random manner.

Please read the information in the appendix and solve the following questions:

(1) Develop a mathematical model to understand the structure and organization mechanics of the network. The structural characteristics of the different types of networks and the organization principle are not always the same.

(2) Propose an effective method to identify the error connections. Show the completeness of how the structural characteristics are discovered; explain the validity and the accuracy of the mathematical model as well as the accuracy of the algorithm.

Attachment

Data description

The networks related to this problem are numbered 1 to 6 in Table 1. The data itself and its detailed description of how it can be obtained are given in the **Supplementary information**.

Table 1： Brief data description

No.	networks	types	nodes	Total connections	Error connections
1	Social network	undirected	50398	44268	2108
2	Social network	directed	25440	1506389	71732
3	Bio-network	undirected	2186	10491	499
4	Bio-network	directed	293	2263	107
5	Info-network	undirected	4554	5788	275
6	Info-network	directed	2591	9093	433

For any of the above network, if the real number of error connections is R, then the player should submit how those R error connections are identified in a standard format (please refer to **Supplementary information** for the standard format of submission). If **r** out of R error connections are identified correctly in the submission, then the score is r/R. The total score obtained by the players in all 6 networks is the only measure of the accuracy of the algorithm.

Supplementary information

1） To get the data, please log in **http://www.pkbigdata.com/common/competition/150.html**, and get the right to download data and submit the results after register with your real name. In order to guarantee fairness of this competition, all teams must register with their real names, and each team can only register once. Anyone(any team) who does not register in real name, or register more than one team names or maliciously affect the registration of other teams, will be disqualified. When registering the **www.pkbigdata.com**, the register email must keep the same as the registration on the **www.saikr.com/apmcm**. Besides, the team name on the pkbigdata.com is your team number, for example 0001, 1100.

2） Please note that each team not only need to submit the final paper, but also must submit the algorithm and the results. In the contest page entitled "identify the error connections in the network", the players can see a more detailed instruction of the data and the submission format of the results. During the contest, for each network, each team has no more than 10 chances to submit the results of the algorithm. Players can obtain a real-time ranking of all the teams for the algorithm accuracy of each network and the total accuracy score of all 6 networks. Please submit the results with the right team name as the one in the registration platform. Although the submission of the algorithm results is not necessary in this contest, the results have a great impact on the final results of the competition.

3） The score defined by r/R need show on the abstract of the final paper. The expert group can consider the score but not all. Novel ideas are more important.

The following two books published by Higher Education Press maybe helpful for the participants to understand the problem and design algorithms:“Network Science: An Introduction”(Wang xiaofan et al., 2012); “Link Prediction” (Lv Linyuan et al., 2013).