Homework 1

Deadline: 2018.04.15 (Sunday) 23:59

* Competition System: http://140.116.247.120:8000/

Problem: Community Detection in a Social Network

Network community detection is one of the essential tasks in social network analysis. In the lecture, you have learned a variety of algorithms to detection communities. In order to help you better understand the structure of communities, in this homework, you are asked to develop your own community detection algorithms using either Python or R. The most special part of HW1 is that it is a competition (some of you might feel excited about competition ☺). Your developed methods will be compared with not only your classmates, but also a set of conventional and state-of-the-art methods of community detection. Note that we do NOT provide you the number of ground-truth communities for this social network. In other words, your task is to not only determine the number of communities, but also which nodes belong to each community.

In the following, we provide you the settings for the competition in this homework.

- **Social Network Dataset.** We provide you a small-scale social network data with 5,000 nodes and around 20,000 edges. You can download the network data in Moodle. The data contains a list of edge, and each edge represents the connection between two nodes' IDs.
- **Competition System with Rules.** To upload your detected communities results, you should follow the instructions of provided by TA to see the format of the uploading file. The uploaded file is he pure text format. In the file, each line is a pair of "node ID" and "community ID" separate by TAB. For example, "17 33" means node 17 belongs to community 33. It is also important to let you know that each team has only 20 times for the submission within a day. So please cherish your submission times and do not intend to perform try-and-error tests that may waste your submission times.
- **Performance Evaluation.** Since we have the ground-truth communities for the network data, we can evaluate any methods by computing some accuracy measures. We take advantage of two commonly used evaluation metrics of clustering: one is *Normalized Mutual Information* (NMI), and the other is *Accuracy in the Number of Communities* (ANC). For NMI, you can refer to the following two URLs for the details.
  https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html
  http://scikit-learn.org/stable/modules/clustering.html#mutual-information-based-scores
  As for ANC, it is the relative accuracy between the detected and the true number of communities, given by: $ANC = 1 - \frac{||C^*| - |\hat{C}||}{2|C^*|}$, where C^* is a set of ground truth communities and $\hat{C}$ is a set of detected communities. It is important to note that NMI takes 75% and ANC

takes 25% of your final grade for this homework.

- **Competing Methods.** The competitors of your methods include a Random clustering, KL algorithm (the oldest method), K-means clustering (the conventional unsupervised mining method), Spectral clustering (the matrix-based method), and Louvain algorithm (the state-of-the-art method). Their NMI and AUC scores have been shown in the competition system.

You are asked to submit your community detection results to the competition system, submit your source codes with clear comments, and describe the details of your methods in a report containing the following justification, and submit the code and the report in Moodle. The maximum length of the report is 10 pages using this template: https://www.acm.org/publications/proceedings-template . Note that you are encouraged to use LaTeX to compile your report and submit your report in PDF.

- **Description.** What are your methods for detecting network communities and determining the number of communities? What are the main ideas, intuitions, and physical meanings of your methods? You are asked to write down the detailed procedures (e.g. algorithms) of your methods. Please also give a title of your report, and name your proposed methods in the report.
- **Analysis.** You need to analyze why your methods lead to high or low accuracy scores by varying some parameters if any. You might want (highly recommended but not necessary) to answer questions like: when does your method work better? For all the methods you have tried, which is better and which is worse and why? Any methods lead to good NMI but bad ANC, or bad NMI but ANC? Which parameters (if any) significantly affect the accuracy of your methods? What about the running time in seconds (time efficiency) of your methods? What are the strong and weak points of your methods? Have you combined the results of several methods to produce the detected communities? If so, how do you make the combination? Note that if none of the methods that you had tried and developed result in worse accuracy, it's fine. Then the grade of your homework will highly depend on both of your description and the analysis in your report. Therefore, it would be better for you to write down the details about all the methods you have tried, show the abovementioned items, and analyze why it cannot the methods you have tried lead to worse results in your report. With your report, we are able to understand which methods cannot work even though they possess some physical meanings.
- **[Optional] Visualization.** Seeing is believing, again. You are asked to plot some figures of the detected communities (by varying some parameters if any), together with some textual description, to explain the effects of your developed methods.
- **[Optional] References.** If you methods are implementing some of existing community detection algorithms searched in Google (Note again that you cannot directly call any community detection functions in any packages written by others, but you can modify and extend them.), you still need to include the description part in your report, and provide the

references to the corresponding papers. If you totally have no ideas about how to detect communities, we have provided you some survey papers for your references in the following. The first paper [1] well reviewed a number of community detection algorithms, the second one [2] empirically compared the performance of different methods, and the last one [3] compared the usages and applicable scenarios for a collection of methods.

References

[1] Santo Fortunato. Community detection in graphs. arXiv:0906.0612, 2009. (6417 cites)

[2] Jure Leskovec, Kevin J. Lang, and Michael Mahoney. Empirical Comparison of Algorithms for Network Community Detection. ACM International Conference on World Wide Web, 631-640, 2010. (756 cites)

[3] Survey and comparison for a collection of community detection algorithms. https://github.com/Lab41/survey-community-detection

How to Submit Your Homework?

You will need to submit multiple files. One is your Python/R code, and the other is the report in PDF format. Please name the source code file as “hw1.py” or “hw1.R”. If you have developed multiple methods, you can name them as “hw1_XXX.py” and “hw1_YYY.py”, where XXX and YYY are the names of your methods. In addition, please also submit your report as “hw1.pdf”. Finally, zip your files and submit the file with file name “姓名_hw1.zip” using Moodle.