Homework 3

Deadline: 2018.05.29 (Tuesday) 23:59

* Competition System: (TBA)

Problem: Influence Maximization in a Social Network

Viral marketing is very important business application based on social networks. To identify the most influential individuals to perform viral marketing, influence maximization, finding a seed set that can lead to the maximum spread of influence, is the key technique in social network analysis. In the lecture, you have learned a variety of algorithms to select the seed set for influence maximization. In order to help you better understand how influence propagation and maximization work for real applications, in this homework, you are asked to develop your own influence maximization algorithms to select the seed set using either Python or R. The most special part of HW3 is that it is a competition (hope you had enjoyed the competition in HW2 and are willing to do it again ☺). Your developed methods will be compared with not only your classmates, but also a set of conventional and state-of-the-art methods of influence maximization. Note that HW3 will have NO so-called ground-truth (i.e., the best seed set) since it is impossible to find the optimal solution as explained in the lecture. In other words, your task is trying to come up with a method that can approximate the influence spread of the greedy algorithm as close as possible. If you can beat the greedy algorithm, the method you proposed will be truly potential to get published in top conferences and journals. :p

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 around 15,000 nodes. You can download the network data in Moodle. The data contains a list of edges, and each edge represents the connection between two nodes' IDs.
- **Competition System with Rules.** To upload your 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 "node ID". For example, "17" means node 17 in a selected seed. 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.** By following the typical setting of influence maximization, we will evaluate the performance of any methods by computing the **influence spread** of a seed set based on the **Independent Cascade** model. The **propagation probability** of each edge in the social network is set as **0.05**. The size of the seed set is **30**, which means that you are asked to **select at most 30 influential seeds**, and submit such **30** seeds to the competition site.

- **Competing Methods.** The competitors of your methods include a Random seed selection, the Degree Discount Heuristic, the Degree Heuristic and the New Greedy IC algorithm (the state-of-the-art method). Their influence spread scores have been shown in the competition system.

You are asked to submit the list of your selected 30 seeds 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 finding the influential seed set in maximizing the spread of influence? 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 influence spread 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? Which parameters (if any) significantly affect the influence spread 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 resulting seed set? If so, how do you make the combination? Note that if NONE of the seed sets generated by your methods that you had tried and developed can lead to high influence spread, 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 may want to visualize and highlight the seed nodes in the social network (by varying some parameters if any) so that we can understand the actual position of influential seeds in a visual form, together with some textual description, to explain the effects of your developed methods.
- **[Optional] References.** If you methods are implementing some of existing influence maximization algorithms searched in Google (Note again that you cannot directly call any influence maximization seed selection 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 find the seed set in influence maxmization, we have provided you some papers for your references in the following. These three papers [1][2[3] have delivered state-of-the-part methods for maximizing the spread of influence in the independent cascade model.

References

[1] Y. Tang, X. Xiao, and Y. Shi. “Influence maximization: Near-optimal time complexity meets practical efficiency.” In ACM SIGMOD, 75-86, 2014. (116 cites)

[2] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck. “Influence maximization: Near-optimal time complexity meets practical efficiency.” In ACM CIKM, 629-638, 2014.

[3] H. T. Nguyen, M. T. Thai, and T. N. Dinh. “Stop-and-Stare: Optimal Sampling Algorithms for Viral Marketing in Billion-scale Networks” In ACM SIGMOD, 695-710, 2016.

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 “hw3.py” or “hw3.R”. If you have developed multiple methods, you can name them as “hw3_XXX.py” and “hw3_YYY.py”, where XXX and YYY are the names of your methods. In addition, please also submit your report as “hw3.pdf”. Finally, zip your files and submit the file with file name “姓名_hw3.zip” using Moodle.