General Exam

Jason Portenoy

Please answer questions 1-4 in approximately 4-5 pages, not including references. For question 5, provide a link to your github/bitbucket repository with a 1-2 pg summary of what you did.

1. **What is community detection?** What are the common algorithms and history? How is this problem approached in various disciplines such as computer science, physics and network science? How does one evaluate a community detection algorithm? What are the top performing algorithms? Please give special attention to the evaluation part of this question.

2. **Applying community detection.** How is community detection being used in the wild? What kinds of questions does it try to answer? How is it being used in different disciplines, industry, etc?

3. **Visualizing communities.** What is the state of art for visualizing networks and specifically communities extracted from these networks? What are papers and labs developing these techniques? What are open challenges for visualizing network communities?

4. **Measuring our confidence in communities.** How does one measure confidence for a given cluster? How much would the partitions change with a different seed for a given algorithm or removal of a link or node in the raw network? To narrow your answer, please focus most of your answer on InfoMap since this is the algorithm you are most familiar with.

5. **Coding community detection algorithms.** To test your coding skills and your knowledge of community detection, recode in python a rudimentary version for minimizing the MapEquation, with a particular focus on the search algorithm. There are open source versions available (e.g., InfoMap) and you can refer to these, but I want you to focus on some of your own search solutions for minimizing the map equation. For example, you could employ a greedy algorithm or Monte Carlo based version. You can run this on a small example network. I do not expect a full scale code base. I just want to see some of your ideas for how you would minimize the MapEquation.

If you have any questions, let me know.
Good luck!

History of community detection

Guide in Fortunato introduction

- Earliest examples: fortunato p.4
- Social sciences: social ties
- computer science: parallel computing. graph partitioning (since 1970s)
- Girvan and Newman paper 2002. Physicists enter the game.

Across many disciplines of science, it is common to encounter data that can best be represented as a network, with entities linked to each other through association, flow, or some form of connection. These entities are represented by *nodes* or *vertices* connected to each other with *links* or *edges*. This overall

representation is known as a *network* or *graph*. The idea of community detection as a research field comes from a basic intuition that there exist in these network data groups of nodes that are structurally more related to each other than they are to members of other groups. Within the community comprising the new, interdisciplinary field of *network science*, the concept of *community* has been somewhat more formalized as a group of nodes (a *subgraph*) with a high concentration of edges connecting vertices within the group, and a low concentration of edges with nodes outside the group [1].

References

[1] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, Feb. 2010, 05400, ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.11.002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0370157309002841 (visited on 04/26/2017).