Complex and Social Networks: Lab session 5

Finding community structure

Miguel Alcon & Egon

Contents

TASK 2

Introduction

In this section, we are going to analyze the resultant communities of applying one community detection algorithm to a huge network, the Wikipedia network. Since this network is a directed graph, only two algorithms can handle it, which are `edge.betweenness.community`, `label.propagation.community` and `walktrap.community` (the other ones crashed). Only the `walktrap.community` algorithm finish its execution in a reasonable time, hence it is the one we used.

Analysis of the communities

The algorithm groups the nodes in 3352 communities. In the following figure, you can see the distribution of the nodes along the communities.
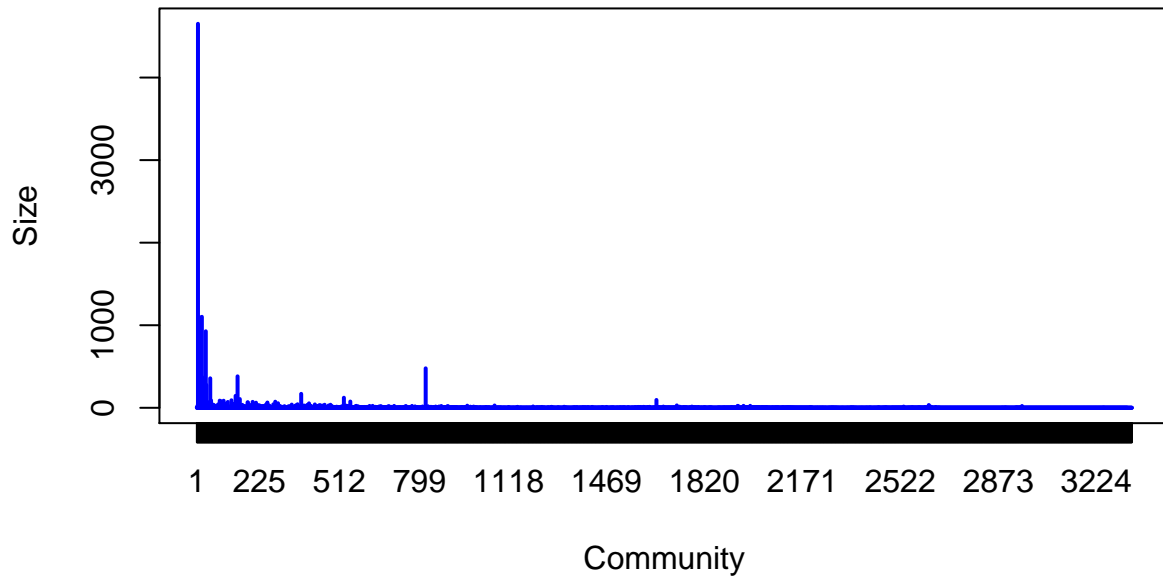
As you can observe, nodes are not very well distributed. We have only 15 communities that have more than 100 nodes, while 1528 of them have less than 3 (only 16 have only one single node). However, we are going to take a look at the labels of the nodes of some of these communities.

We selected randomly 10 communities:

1. "Kd-trie", "Prefix hash tree", "Trie", "Burstsort", "Hash trie" and "Acyclic deterministic finite automaton".

2. "Etherloop" and "High Speed Voice and Data Link".

3. "Fleet Satellite Communications System" and "UHF Follow-On System".

4. "Loop Maintenance Operations System" and "Craft access system".

5. "Sheikh Ahmad Dede" and "Heaven & Earth (TV series)".

6. "Comedy of errors", "Havoc, Inc", "Farce" and "Body swap".

7. "Family Ties (Farscape episode)", "Meltdown (Farscape episode)" and "A Human Reaction".

8. "Penrose method", "One man, one vote" and "Plural voting".

9. "Porto Alegre Manifesto" and "Armand Mattelart".

10. "Elizabeth Moberly" and "Environmental factor".

Doing a little bit of research in Wikipedia, we found a common point between nodes of each community:

1. Tries.

2. Voice and data transmision.

3. Satellite communication systems of the U.S. Navy.

4. -

5. TV show and a man who appeared in.

6. Literature.

7. Episodes of the Farscape TV show.

8. Voting.

9. A manifesto and a man who signed it.

10. Cancer.

In point 4, we could not find any entrance in Wikipedia for "Craft access system". Moreover, in point 10, we have "Elizabeth Moberly": a British research psychologist and theologian who became involved in cancer research, and "Environmental factor", where the main result of it is cancer. The relation is almost inexistent.

We also had a look to the largest community (4651 nodes) to see if there is a relation between its nodes. We selected randomly 10 elements of the community, which are: "Bairstow's method", "Boolean algebra (structure)", "Ideal (set theory)", "Pappus graph", "Parity (mathematics)", "Semi-continuity", "Elongated triangular cupola", "Methods of contour integration", "Jensen's Device", "Domain (ring theory)". Again, doing some research in Wikipedia, we notice that all nodes are related with mathematics.

Conclusions

To sum up, community detection algorithms (`walktrap.community` to be concrete) group nodes into communities correctly, but probably not the best way. I.e., the communities it creates are good because nodes within it are related, but they could be inside one where they fit better. Furthermore, we are sure that most of the small communities can be in a bigger one with a more general topic, for instance, the TV shows of points 5 and 7.