

open.michigan

Unless otherwise noted, the content of this course material is licensed under a Creative Commons Attribution 3.0 License.

<http://creativecommons.org/licenses/by/3.0/>

Copyright 2008, Lada Adamic

You assume all responsibility for use and potential liability associated with any use of the material. Material contains copyrighted content, used in accordance with U.S. law. Copyright holders of content included in this material should contact open.michigan@umich.edu with any questions, corrections, or clarifications regarding the use of content. The Regents of the University of Michigan do not license the use of third party content posted to this site unless such a license is specifically granted in connection with particular content objects. Users of content are responsible for their compliance with applicable law. Mention of specific products in this recording solely represents the opinion of the speaker and does not represent an endorsement by the University of Michigan. For more information about how to cite these materials visit <http://michigan.educommons.net/about/terms-of-use>.

 UNIVERSITY OF MICHIGAN



PROBLEM SET 5 – Week 7

1. The Girvan-Newman betweenness clustering algorithm (50pts)

Select a network of up to 200 nodes, preferably one where you suspect there might be interesting structure. For 90% credit (45pts), you can use the `poliblogrecip.gdf` file, which I had demoed in lecture, but for max credit, you need to construct a network of your own (based on real data, or an artificial topology of your choosing). Your network should be undirected. Load the data file into Guess and run the script `betweennessclustering.py`. The GUESS tool has a button, which will remove the edge with the highest betweenness (marked red) at each click, and another button to keep removing edges until the next community is broken off (or all edges are removed). You need to decide when to stop removing edges (fyi, this is just an educational tool implementing a slow algorithm and it cannot cluster big data sets). Please answer the following:

- Does the algorithm allow you to identify underlying communities in the network?
- Is the removal of a leaf node a good stopping criterion for the algorithm? That is, if you stop removing edges before a node with degree one is removed, does this reveal informative community structure, or would you have wanted to stop earlier (or keep going after removing a leaf node).
- Which nodes are not very embedded in their communities? How does the algorithm reveal this fact?
- Are any nodes 'misclassified' by being placed in a cluster where you think they may not belong according to a node attribute? How can such a misclassification be informative?
- Turn in one image that shows community structure in this network *I*.

2. Hierarchical clustering (50 pts)

Use Pajek to hierarchically cluster the very same network (again 45 pts at maximum if you use the political blog network, 50pts if you use your own network).

- Visualize the permuted adjacency matrix (*I*)
- Visualize the dendrogram. (*I*)
- Compare the hierarchical clustering with the betweenness-based clustering at a point where both methods produce the same number of communities (e.g. both have 2 communities). Are there any differences in the groupings? Which nodes were classified in different communities? Use your knowledge

of what the network represents to discuss why some nodes were placed in different communities by the two different techniques.