### Analysis 8 Overlapping Community Detection - Master

### Construct your solution using Analysis-8-Overlapping-Communities-Lastname.Rmd

### (You will do some of the work in Gephi, but include screen dumps in the R Markdown document.)

### 

### Due Monday April 25 11:55 pm but accepted until Laulima shuts down.

### 

### 68 points possible

### Introduction

This is a follow-up to the community detection assignment, in which we look at overlapping community detection. We will work with the Network Science network. Although this is a weighted graph, for the purposes of the assignment **link communities should be computed UNweighted**. There are two reasons for this:

* The weights were constructed by Newman to model strength of collaboration, as reflected in co-authorship possibly across multiple papers. However, the questions I ask below are only concerned with dichotomous presence or absence of co-authorship, not strength of collaboration. So an unweighted network better matches the questions at hand.
* Leaving weights out of it will simplify computation of link similarities in question 2 and your other explanations, as you need only examine the presence or absence of links and community membership. (I ran this assignment with weights and it gets messier.)

### 1. Computing and Visualizing Link Communities (14 pts)

The following is a slight variation on the essential steps we did in class. It is included to be sure you can compute and visualize link communities, and so you have the visualization to inspect for the following questions. Follow the example of 12-6-reify\_link\_communities-Demo.R, but do not include unnecessary steps: we will take off points if you seem to be just re-running the class demo without understanding which parts are needed.

**a. Preparation of the Network (4)**

1. Read in and summarize netscience.graphml
2. Compute **unweighted degree**, **weighted degree**, **unweighted** **page rank**, and **unweighted** **betweenness** centrality for the vertices in this graph, and assign them to vertices as attributes. (2)
3. Compute **unweighted cluster\_louvain** and **unweighted cluster\_infomap** community membership, and assign these as attributes to the vertices. (We want unweighted so we can compare to the link communities, which are unweighted for reasons given above.) (2)

**b. Computing Link Communities (6)**

1. Compute **link communities**. Use the **unweighted** edge list, default hcmethod "average". For reasons that will be obvious if you do otherwise and try to preview the document, you should turn off verbosity and plotting in the call to getLinkCommunities. (DO NOT USE `{r echo=FALSE}: this will make the code invisible in the Knit html.) When done, print the link community object (with 'print') to show the summary. (3)
2. Use **reify\_link\_communities** to create a copy of the graph with link communities reified as nodes. Summarize the new graph. (2)
3. Write this copy out as graphml. Include this in your homework submission in case we need to inspect it. (1)

**c. Visualizing the Link Communities (4)**

Read the modified graph into Gephi. Visualize the graph that you just loaded into Gephi in a manner that makes the link community structure visible. The link community nodes should be a different color than the actor nodes (I recommend yellow to make them salient). The layout algorithm should place related actors and link communities together on the screen, but no overlap between nodes. Actor and community labels should be visible. (2)

**Use the Preview facility to generate PDF** for the visualization so that it is readable even when zooming in: we may need to refer to details of your visualization. (2)

Keep Gephi running so that you can inspect it and make bitmap screen dumps to embed in your response below.

### 2. Basics of the Link Communities Algorithm (17 pts)

In the first two questions you will explain, in terms of how the algorithm works, why there are some nodes that are not classified into communities. (These nodes are probably on the periphery of your layout.) Refer to Barabasi's Figure 9.23, or Ahn et al’s Nature (2010) paper for full details, and the igraph documentation of getLinkCommunities for reminders of how link communities are computed and used to induce node communities. In the third question you explain the computation of link communities.

*Procedural Advice:* To find nodes by name: go into the Data Laboratory, sort by Label; right click on desired node and select "Select on Overview"; and go back to Overview. (You may need to click the arrow and then the hand icon to get manipulation control back.)

**a.** (1) Why are no isolate nodes (e.g., "FREEMAN, L") classified into node communities?

**b.** (2) Why are no isolated dyads (e.g., "MA, H" and "ZENG, A") or 2-paths (e.g., "CSETE, M", "DOYLE, J", and "CARLSON, J") classified into node communities, even though they have links between them? LIKE MOST STUDENTS LAST YEAR, YOU WILL GET THIS WRONG IF YOU GUESS. LOOK AT THE DOCUMENTATION, which provides half the answer. A plausible assumption is needed for the other half.

**c.** (14) Explain the three communities associated with "GEISEL, T", in terms of how link similarity is computed, as directed below. *Procedure:* First rearrange the nodes in the visualization of Geisel’s ego network so the structure is clearer. Then make a screen dump and include it as a figure in the .Rmd for reference (2 pts). Then answer in four parts:

* **Identify the link communities** by name and listing the links in each of the link partitions. Use notation {GEISEL, DODEL} etc. to identify links (edges, not vertices). (3)
* **Perform similarity metric computations** (Barabasi Figure 9.23) on pairs of links. You can rely on isomorphism to avoid repeating work. Specifically, there are three situations:
  + pairs of edges that are within the same link community and share Geisel (e.g., {Geisel, Timme}, {Geisel, Wolf}),
  + pairs of edges that are within the same link community but do not share Geisel (e.g., {Timme, Wolf}, {Timme, Geisel}), and
  + pairs of edges that are in different link communities but share Geisel (e.g., {Timme, Geisel}, {Herrmann, Geisel}).

(We do not compute similarity between edges that do not share a node.) **Your answer should be three numbers, the similarities for each of these situations.** (3)

* **Discuss the specific algorithmic steps by which these results of similarity computations were used to produce the above link communities.** (We *already* computed the similarities above: don’t tell me how they are computed again!!!) Which of the above links of which similarity were merged first? What happens next? What is the role of the density function? Do not give a generic description of the algorithm; describe the specific handling of the above similarities. (3)
* Then **identify the node communities** that are induced on nodes from the link communities by listing the nodes (not links) in each node community. (3)

### 3. Link Communities versus Ego Networks (15 pts)

These two questions show two ways that link communities differ from ego networks, by inclusion and by exclusion.

*Procedural Advice:* You can quickly identify and compare link community membership and co-authorship ego networks by mousing over the link community and author nodes in Gephi, watching how highlighted nodes change.

**a.** (3) Watts is in the same induced node community as Hopcroft and Kleinberg, but there is no direct link between them, indicating that Watts did not publish with either of them (as of this data capture). Explain in terms of the algorithm and network structure how they got into the same induced node community.

**b.** (12) In this data, Newman publishes with (has links to) four authors who are NOT in his link (induced node) community, but are in other link communities.

* Make a supporting figure that includes both Newman’s ego network and his link community. Easy way: Filter for Newman’s ego network of order 2.
* Identify these 4 authors, grouping them in two groups of two due to their distinct structural situations (which can be explained as part of the next item).
* Explain why these authors are not in Newman’s link community (COM\_305). Due to their distinct structural situations, you will need to give one explanation for two of the linked authors, and a different explanation for the other two.
* What other link communities (as we computed them), if any, are the four **links** from Newman to these authors in? Why is that the case? Note I am asking about **links**, not nodes. (2)
* Are the ties that Newman has to these authors (those that he has published with but are not in his link communities) strong or weak ties (as defined by Granovetter and discussed in Easley and Kleinberg readings)? On what evidence do you draw this conclusion? (2)
* What does this have to do with structural holes and Newman's role in the field? (2)

### 4. Interpreting Induced Overlapping Node Communities (16 pts)

While other questions in the assignment are about your understanding of the method, this question is where you show how link community configurations correspond to the domain being modeled. In other words, this is the question that shows you what link communities are good for! We start with a simple one.

**a.** (4) Barabasi and Albert published influential papers together (the BA model).

* Identify the link community that they both belong to and show the membership of the induced node community (type in the community ID and the names of the authors). There will be other authors present. (2)
* For an additional point, include a figure showing the link community (filter for ego network of the COMM node), in which case you do not need to type in all the names. (1)
* Use Google or Google Scholar to find a paper that has as many of these co-authors as you can (all but one is OK) to show why they belong together. Give the citation of the paper in bibliographic form. (1)

**b.** (12) Another prominent author is "PASTORSATORRAS, R" (Pastor-Satorras), who is involved in edges in several link communities. Use Google Scholar to figure out which link community is associated with publications on each of the following three topics (which, by the way, show the diversity of network science applications):

* Epidemic spreading in networks
* Protein interaction networks
* Social distance attachment

As above, give the link community number for each and the induced node membership of each link community in your data, optionally include a figure for an additional point, and provide a representative publication involving as many of the authors as you can.

*In addition to showing the diversity of his work, this is an example of an analysis that would not be possible if Pastor-Satorras were placed in one partition by a conventional community detection algorithm. If you find his work interesting, see his* [*Google Scholar page for recent work*](https://scholar.google.com/citations?hl=en&user=Cd8kFioAAAAJ&view_op=list_works&sortby=pubdate)*.*

### 5. Comparing Link Communities to Partition Methods (6 pts)

We know from prior work that InfoMap usually gives a finer partition than Louvain. Here we will compare Lovain and InfoMap to link communities. (Do not compare Lovain and InfoMap to each other: we did that last week.)

*Procedural Advice:* Color nodes by the Louvain or InfoMap partition, but select your community node color (e.g., yellow) for the "null" value.

* We already know that InfoMap usually provides finer granularity partitions than Louvain. How would you characterize the relationship between the granularity of overlapping node communities induced by link communities and the Louvain and InfoMap partitions? (2)
* Obviously Link Communities provide overlapping communities while Louvain and InfoMap force partitions. What other structural feature does Link Communities find? Use an example from the visualization. (2)
* Provide screen dumps of the main part of your graph that supports your conclusions (zoom in to make it visible) showing the partitioning by Louvain, and then by InfoMap, both with the link community nodes shown. (2)

### For the curious

In previous years, this assignment asked how the link communities change if we change hcmethod to "single" or "complete". We found that the number of link communities changes. Then we interpreted an apparent discrepancy between the hcmethods for the connected component for "WASSERMAN, S" and "FAUST, K", who published an important book on social network analysis together. In two hcmethods, Faust and Wasserman are each in their own induced node communities and Pattison is not classified. Under a third hcmethod Faust and Pattison are in Wasserman's induced node community but Wasserman is not in Faust's. This required some thought.

I also tried removetrivial=FALSE and use.all.edges=TRUE. The former did not add many communities, because most triplets of connected nodes are triangles, not just two-paths. The latter was a lot slower and gave more communities, although I have not yet found links that do not share a node that ended up in the same community.