### Analysis 5: Centrality Metrics and Local Structure

## Construct your solution using Lastname-Analysis-5-Local-Structure.Rmd

### Final version due in Laulima evening of Saturday March 26. 80 points available.

### Part A: Problems on Centrality Metrics

You will:

* show that you know how to compute the metrics in both igraph and Gephi;
* discuss discrepancies between igraph and Gephi results;
* compare the metric results to each other and explain the differences;
* and draw some conclusions about the Network Science domain.

The actual work in igraph and Gephi (questions 1 and 2) is trivial and follows the class demonstrations. The harder part is thinking about and explaining what is happening (questions 3 and 4), but even that should be easy if you studied the class demonstrations well. Question 5 requires the most original analysis and discussion.

A substantial amount of work is done in Gephi as well as R, but we will consolidate the results in an R Markdown document since problems in the next section also use R. In this section, you use R to compute igraph versions of the vertex centralities and annotate the vertices with the results. You then write this out as .graphml, read into Gephi, compute more metrics in Gephi, and do the remaining work using Data Laboratory and the Overview visualization.

## 1. Computing and Inspecting Centralities in igraph (6 pts)

**a.** Load netscience.graphml (weighted undirected) into igraph as "NS", compute the following centrality metrics, and store the results as attributes on the vertices using the indicated attribute name, for example, V(NS)$i\_degree <- degree(NS). (We use i\_ to distinguish those computed in igraph from those in Gephi.) Please do it in the order indicated so that the grader can quickly compare results as they will later show up in Gephi. (6)

* V(NS)$i\_degree: Unweighted degree
* V(NS)$i\_strength: Weighted degree
* V(NS)$i\_eigen\_cent\_u: Unweighted Eigenvector Centrality
* V(NS)$i\_eigen\_cent\_w: Weighted Eigenvector Centrality
* V(NS)$i\_page\_rank\_u: Unweighted Page Rank
* V(NS)$i\_page\_rank\_w: Weighted Page Rank
* E(NS)$distance: Edge weights adjusted by the negation method (d = max(w) - w + 1; see Adjusting-Weights-in-Network-Science.R) to be distances for path-based centralities. Uncomment the histogram code to display a histogram of your result.
* V(NS)$i\_between\_u: Unweighted Betweenness centrality
* V(NS)$i\_between\_w: Weighted Betweenness centrality using $distance
* V(NS)$i\_closeness\_u: Unweighted Closeness centrality
* V(NS)$i\_closeness\_w: Weighted Closeness centrality using $distance

Normalize all metrics where you have that choice above as well as in Gephi (problem 2) so they are comparable. We are computing unweighted and weighted versions in igraph where applicable so we can compare to Gephi's results. The path based metrics require the adjusted $distance weights so that values can be interpreted as distances.

*Comment:* Hubs and Authorities *can* be computed on undirected graphs, but they will be identical to each other since there is no distinction between in- and out-degree. I used to include Hubs in this assignment, but the conclusion is simple: in an undirected graph it is identical to Eigenvector centrality. You can see why by looking at the formulas in Newman.

***Show that you have successfully annotated the graph*** **with**

* summary(NS)

NOTE: If you get results like "i\_eigen\_cent\_u (v/x)" (notice the x), and/or inspection of values shows nested structure like

> head(V(NS)$i\_eigen\_cent\_u)

[[1]]

[1] 5.451790e-17 8.235003e-17 9.933719e-17 9.752832e-17 1.032684e-16 …

and/or you cannot write out the attribute to graphml, then you probably had a previous error “number of items to replace is not a multiple of replacement length” that created the attribute with the wrong type. Delete the attribute and try again using $vector. A proper result should show up as "i\_hub (v/n)" and inspection of values look like:

> head(V(NS)$i\_eigen\_cent\_u)

[1] 6.788641e-17 1.195143e-16 …

## 2. Computing and Visualizing Centralities in Gephi (6 pts)

**The script will write out the resulting graph as netscience-with-metrics.graphml. Include the .graphml in your submission.**

**a.** Load this annotated graph into Gephi and ***compute the following node centrality metrics*** using Gephi's methods ***with adjustments***, namely:

* Eigenvector: increase iterations to 1000
* Page Rank, unweighted and decrease epsilon to 0.0001 or 1.0E-5 1.0E-5
  + (If you wanted to do weighted as well, you'd have to save the prior weights, copy the igraph distance to weight, and copy the unweighted vertex PageRank results before running the weighted version. I decided to not require this.)
* Betweenness and Closeness: normalized (read the documentation on the dialog).

**b.** Visualize the graph in a manner that will make the clusters and labels clear. I suggest:

* Scale node size by Degree (you will change this later).
* Run Connected Components and color the nodes by component ID (top 30 or so OK).
* Run the OpenOrd layout to separate the clusters quickly
* Follow up with a few seconds of ForceAtlas 2, Gravity 5, and Prevent Overlap to clean it up.
* Turn on labels and scale them to fit.

Note: When you change node size for other metrics you'll need to re-run ForceAtlas 2 for a second or two to remove overlap.

***==>* Save your Gephi project as Lastname-Analysis-5-Centrality-Metrics.gephi to be included in your homework solution.**

You do not need to make a screen dump or PDF of the visualization, as we can look at your Gephi project.

### Another Prelude

In the next two questions you will compare different centralities for the nodes, and identify, explain, and interpret any differences found. You can do this comparison by sorting the rows in the Data Laboratory by the two centralities being compared and seeing whether they give the same results. Additionally, resizing the nodes by the centralities being compared in Overview and noticing what changes helps you see the structural settings of the changes.

*Suggestion:* To set up for questions 3 and 4, go into Data Laboratory and use the light bulb with a + icon, upper right of display, to uncheck (remove from the display) any metrics or columns not being used (Id, Interval, Authority, Eccentricity, Harmonic Closeness Centrality). Then check the columns that we are using to make them visible. Then, as you answer the questions, sort the columns to put the Gephi and igraph versions of the metrics next to each other (e.g., i\_eigen\_cent\_w/u next to Eigenvector Centrality) so they are easy to compare. (Unfortunately the ordering reverts to default when you check or uncheck columns.)

I recommend that you use both the visual (Overview) and tabular (Data Laboratory) representations for each comparison. The tabular representation helps you find differences, and the visual one puts those differences in context. For example, *are the nodes involved in a small component or the large component? Do the nodes with different values have any particular connectivity pattern?* You can move between the two representations by right-clicking on the item of interest and using “Select …” before moving to the other representation.

***Your explanations should rely on and refer to your knowledge of the following, where applicable:***

* ***the mathematics of the metrics*** (e.g., as discussed by Newman and summarized in my lecture);
* ***alternative parameters and implementation strategies***, as discussed by Newman and documented in the software (e.g., handing of weights and of disconnected graphs); and
* ***the domain being modeled*** (e.g., knowing the names of important authors in Network Science; thinking about how isolated versus well known researchers and groups might function).

***Please note:*** the above three items are often neglected by students who are only reacting to the superficial results. These are important for understanding the meaning of what you are doing.

***Be Precise and Informative.*** Don't talk vaguely about the metrics measuring the "importance" of nodes. There are different kinds of importance emphasized by each metric. These may correspond to different social roles. Also, discuss in terms clients can understand. "Node 79" refers to an arbitrary number assigned by the software. "Newman" carries a lot more meaning!

### 3. Comparing Gephi and igraph results (18 pts)

Compare the results for these metrics. **We are most interested in differences between Gephi vs. igraph**, although you might also comment on igraph’s weighted vs. unweighted where applicable. Please follow the HOW / WHY / WHICH / ADDITIONAL COMMENTS format. Reference nodes by their Labels (author names), not by numeric ID. Screen dumps showing the differences are optional: include them only if it is easier to describe what is happening using them, but don't include them if you don't discuss them.

**a. Gephi's and igraph's eigenvector centralities**

*HOW do they differ? Describe how values differ for nodes in this particular graph, particularly comparing nodes in different structural settin*gs.

*WHY do they differ? Refer to documentation on how the metrics are computed and Newman's discussion of alternative approaches, and explain how the approach taken can account for these differences.*

*WHICH do you trust more, and why? Or, if each has their merits, what kind of question would you use each for?*

*ADDITIONAL COMMENTS (anything you wish to add):*

**b. Gephi's and igraph's closeness centralities**

*HOW do they differ? Describe how values differ for nodes in this particular graph, particularly comparing nodes in different structural settin*gs.

*WHY do they differ? Refer to documentation on how the metrics are computed and Newman's discussion of alternative approaches, and explain how the approach taken can account for these differences.*

*WHICH do you trust more, and why? Or, if each has their merits, what kind of question would you use each for?*

*ADDITIONAL COMMENTS (anything you wish to add):*

**c. Gephi's and igraph's betweenness centralities**

*HOW do they differ? Describe how values differ for nodes in this particular graph, particularly comparing nodes in different structural settin*gs.

*WHY do they differ? Refer to documentation on how the metrics are computed and Newman's discussion of alternative approaches, and explain how the approach taken can account for these differences.*

*WHICH do you trust more, and why? Or, if each has their merits, what kind of question would you use each for?*

*ADDITIONAL COMMENTS (anything you wish to add):*

### 4. Comparing the Centrality Metrics to each other (20 pts)

For the following, we will use igraph's versions for consistency. Compare "apples to apples": use normalized values where applicable, and don't mix weighted and unweighted, except in question (a) where asked to do so. Reference nodes by their Labels (author names), not by numeric ID. Screen dumps showing the differences are optional: include them only if it is easier to describe what is happening using them, but don't include them if you don't discuss them.

**a. How do Degree and Weighted Degree compare to each other?** Identify differences in the values or relative ranking of specific nodes as examples. **How do you explain this result** in terms of the metric mathematics and/or what you know about the domain being modeled (network science co-authorship)?

**From here onwards let's use the unweighted versions to simplify our explanations.**

**b. How do Degree and Eigenvector centrality (both unweighted) compare to each other?** Identify differences in the values or relative ranking of specific nodes as examples. **How do you explain this result** in terms of the metric mathematics and/or what you know about the domain being modeled (network science co-authorship)?

**c. How do Eigenvector centrality and Page Rank (both unweighted) compare to each other?** Identify differences in the values or relative ranking of specific nodes as examples. **How do you explain this result** in terms of the metric mathematics and/or what you know about the domain being modeled (network science co-authorship)?

**d. How do Page Rank and Betweenness centralities (both unweighted) compare to each other?** Identify differences in the values or relative ranking of specific nodes as examples. **How do you explain this result** in terms of the metric mathematics and/or what you know about the domain being modeled (network science co-authorship)?

### 5. Conclusions about Network Science Authors (6 pts)

Discuss what the centrality metrics show about the ***roles of the most prominent researchers*** in this domain. Discuss the roles of at least 5 researchers and minimally make use of the following metrics as evidence. Notice that some are weighted, so rankings may differ from what you computed for question 4:

* Degree
* Weighted degree
* Weighted page rank
* Weighted betweenness (using weights adjusted to be distances)

As part of your response, display a table of top 10 nodes under each of the above metrics. Your discussion can reference points made earlier: just be sure they still apply with the weighted versions.

### Problems on Reciprocity, Transitivity and Subgraph Extraction

Two problems involving reciprocity and transitivity follow. In the first, we interpret reciprocity and transitivity results in terms of the processes they imply in the domain being modeled. In the second, we use these metrics as a data integrity check, uncovering a problem that we visualize by extracting a subgraph.

## 6. High Energy Physics Theory Citations (12 pts)

The high energy physics theory citation graph **cit-HepTh.gml** (in folder HEP-Theory-Citation and copied to the assignment folder) is from the e-print arXiv and covers all the citations within a dataset of 27,770 papers with 352,807 edges for the period from January 1993 to April 2003 (it begins within a few months of the inception of the arXiv). If a paper i cites paper j, the graph contains a directed edge from i to j. If a paper cites, or is cited by, a paper outside the dataset, the graph does not contain any information about this.

In this problem we compute reciprocity and transitivity for this directed graph, and interpret the results in terms of domain processes. We need to consider whether the values differ much from expected at random, and whether the values are a natural consequence of the degree sequence or due at least partially to an independent process. To enable these comparisons you will construct G(n,m) and Configuration models for the graphs and compute reciprocity and transitivity on them as well. The grading will be based on your interpretations, not on just showing you can do the computations.

### (a) Generate Random and Configuration Models for Comparison (3 pts)

* Generate a corresponding G(n,m) model with **sample\_gnm** and **directed=TRUE**.
* Generate a configuration model with **sample\_degseq**, specifying both **out.deg** and **in.deg** and using method **"simple"**.

(Notes: Method "simple.no.multiple" did not return on this graph and method "vl" makes undirected graphs. We could also use rewire and keeping\_degseq, but my trials show that the number of iterations needed to sufficiently randomize these graphs take a very long time.)

### (b) Compute Reciprocity and Transitivity (3 pts)

* Compute the **reciprocity** and global **transitivity** for each network being studied.
* Display the results in a nice table (using tibbles or a data frame) in this format:

Graph reciprocity transitivity

HEP #.######### #.##########

G(n,m) #.######### #.##########

Config #.######### #.##########

### (c) Interpret Reciprocity and Transitivity (6 pts)

* **Compare the reciprocity of the given network to those of G(n,m) and the Configuration Model** to see whether the value appears to differ from what is expected at random, and if so whether the value is due to the degree distribution rather than an independent process or constraint. Then, in light of these comparisons, **explain why the result you got makes sense in terms of the real world network being modeled.**

To the extent that a value of the natural network is similar to that of the random model, perhaps randomness explains it. To the extent that a value of the natural network is different from the random model but similar to the configuration model, then whatever processes lead to the degree distribution may explain it. To the extent that the value is different from both models, some process independent of degree distribution must explain it. In either case, describe the relevant domain processes.

* **Repeat the above comparisons and discussion for transitivity.**

Be sure to discuss *both* reciprocity and transitivity (they are 3 points each). Keep in mind that reciprocity is directed but transitivity is computed as undirected in igraph, even if the underlying network is directed.

## 7. Comic Hero Network (12 pts)

**comic-hero-network.graphml** (in folder Comic-Hero and copied to the assignment folder) is derived from the Marvel Social Network Networks of super heroes constructed by Cesc Rosselló, Ricardo Alberich, and Joe Miro from the University of the Balearic Islands (obtained from http://exposedata.com/marvel/ but this site is no longer available). If you need background on this comic book series, see https://en.wikipedia.org/wiki/Marvel\_Comics, although our interpretation only requires understanding the general idea that comic heroes, other protagonists and minor characters appear (and reappear) in comic book issues.

Rosselló et al. offered us a directed bipartite graph, connecting comic book heroes in one partition to the comic issues they appeared in in another partition. So, the out-degree is the distribution of issues heroes have appeared in, and the in-degree is the distribution of how many heroes are in an issue. (Do NOT use the Hero-Coappearance-Network, which is the projection onto Heroes.)

This problem provides an example of the use of metrics to check data integrity and "clean up" errors. (We won’t be comparing this one to random models.)

### (a) Predict Reciprocity and Transitivity (2 pts)

What should reciprocity and transitivity be in a directed bipartite graph as described above?

### (b) Compute Reciprocity and Transitivity and interpret (2 pts)

Compute and display the reciprocity and global transitivity.

Which value(s) are as expected and which are not as expected?

### (c) Identify the Error through Subgraph Visualization (8 pts)

Use the following procedure to find the error (I’m being vague about the metric because you need to answer the above question to identify it).

* Compute the appropriate metric for all the vertices. (1)
* Make an induced\_subgraph of exactly those vertices that have erroneous values. (2)
* Plot the induced subgraph, with vertices labeled by $id and colored according to their bipartite $type, and vertex and arrowhead sizes adjusted for visibility. (4)
* Using the visualization, identify the error and explain how the visualization shows the error. (3)

(If you really want to investigate, I suggest going to Gephi at this point, but decided to end the assignment here.)