### Spring 2022 Analysis 2 Graph Metric Analysis

### Final submission to be uploaded to Laulima before class Thursday Feb 10.

### Up to 50 points possible.

### Prelude

This assignment includes computation of the main network metrics that we studied this week. The assignment is your opportunity to show me that you understand the basics of the most important metrics. **We do this through the device of pretending you are explaining it to a completely naive client.** It also asks you to make some comparisons between Gephi and igraph results, to make sure we understand our tools.

***Use the Simplified EuroSiS WebAtlas Network*** *that is provided****:*** You worked with EuroSiS-WebAtlas.gexf in the previous two assignments. This is a "WebAtlas" constructed for the European Science in Society program. It consists of a partial sampling of Science in Society websites in Europe, and how they link to each other (oddly, the graph is undirected).

In this assignment, *it is critical that you use the simple graph version of this network (not the multigraph you previously used),* because it affects the metrics. To ensure that everyone starts with the correct network I am providing you with EuroSiS-WebAtlas-Simplified.graphml and including the code to load it in the .Rmd template.. The simple graph should be undirected and weighted with 1285 vertices and 6462 edges.

***Using R Markdown:*** You will construct your response in an R Markdown template as in the last assignment. Since some of the values are computed in Gephi, the R Markdown template will tell you where to copy the values from Gephi.

***See Video-210206-R-Markdown-for-Analysis-2.mp4*** in the Videos directory for an overview of how to use R Markdown, Knit to HTML, and zip up your results to upload to Laulima. This video was recorded for last year's class. One difference: Do not remove the Networks folder. We want to be in the habit of creating fully portable and reproducible analyses.

### Basic Graph Metrics in Gephi and igraph

In each of the following questions you will compute metrics in BOTH igraph and Gephi (when available) and compare the results. The points are for execution, interpretation, and explanation and resolution of any discrepancies between igraph and Gephi. This helps build your skills in (a) checking rather than just trusting results, (b) dealing with discrepancies, and (c) communicating results to clients or laypersons. Do not assume that anything is obvious: explain it!

### 1. Degree and Weighted Degree (15 pts)

**(a)** Find the average (mean) degree and average weighted degree of the EuroSiS-WebAtlas-Simplified graph in igraph and Gephi and check whether they are the same. Show the results in the .Rmd table.

**(b)** Explain, in terms of the domain of Science in Society web sites, why it makes sense that average weighted degree would be slightly higher than average degree. (Don't just say "because some of the edge weights are higher": provide your client with an interpretation of what this means in terms of their application area. Look at the actual weights and interpret their meaning in terms of web sites.)

**(c)** Plot the degree distribution of the graph in igraph, using the domain correction and giving it nice labels. (If your igraph plot is a mess, you probably plotted degree instead of degree distribution!) For this network, a lin-lin plot with zero probability values removed will be clearest.

**(d)** Discuss the degree distribution plot as if you are explaining to a client what the degree distribution plot tells us. Explain the *specific* plot, not just how these plots work in general. Relate the features of the degree distribution plot to the application domain (EuroSiS web sites) so your client can understand!

### 2. Distances (15 pts)

*Comment:* "Average path length" (what Gephi calls it) is not as accurate as "mean distance", because we are not averaging over *all* paths, only over the shortest paths (geodesic distances) between each pair of vertices. I use the technical terminology.

**(a)** Find the mean geodesic distance and diameter of the graph in both Gephi and igraph. Use default values in igraph (do not change the optional parameters). (For now we will ignore the centrality metrics that Gephi gives you.)

**(b)** Using the .Rmd table that shows the results, comment on whether the results are consistent across Gephi and igraph. There will be a discrepancy: explain why it occurs and show that you are right by changing something in your igraph call to make the results the same as Gephi.

**(c)** Plot the distances in igraph, giving the plot nice labels. There is no 'distance\_distribution' counterpart to degree\_distribution. There is a distance\_table method, but you need to figure out what $part of the result to plot. (The really large values are because we are counting all pairs shortest paths, and there are vcount(WA)\*(vcount(WA)-1)/2 = 824970 possible pairs.)

**(d)** How would you describe the difference in general shape and range of values between the distances plot of this question and the degree distribution plot of question 1c? (We will return to the significance of this difference in the week on the scale free property.)

### 3. Clustering and Transitivity (7 pts)

Here I use the different terminology of Gephi and igraph for clustering and transitivity because these terms reflect a difference in what is being computed. When we get different results it is important to understand why and resolve the discrepancy in favor of what we intended to compute.

**(a)** Find the "Avg. Clustering Coefficient" in Gephi, the average transitivity in igraph, and the global transitivity in igraph (embedding the R code in your document).

**(b)** Using the .Rmd table that shows the results, compare the three results to each other and discuss why some are the same and some different. Can you argue why the *direction* of the difference between the two results in igraph is what we expect in terms of network structure and how they are computed? (Explain why one is bigger than the other in that particular direction.) *Hint:* See the end of Newman section 7.3, but go one step further than Newman: why this *direction* of difference?

### 4. Components (13 pts)

**(a)** In Gephi:

* Find the number of connected components.
* Then find and identify the three smallest isolated components, which are web sites not connected to the rest of the European Science in Society collection. For each of these isolated components, identify what they are about in terms of country and actor type. *Hint:* Use an Attributes / Partition / Component ID filter, and then look at the nodes in Data Laboratory.)

**(b)** In igraph:

* Write an expression that returns the number of connected components, and nothing else. (The result should be the same as in Gephi.)
* Write an expression that returns the sizes of the components (one integer for each component. Again, it should not return anything else: just the sizes.
* Print out the labels ($label) of the vertices that are in each of the three smallest components. These should be the same as the vertices you found in (a). *Hint*: It's OK to do this with three expressions. V(WA)$label prints out *all* the labels: add a logical test inside square brackets to select the right ones: V(WA)[…]$label

### Comment

If you are curious, repeat the above with the multigraph version of EuroSiS-WebAtlas.gexf (without combining edges). Which metrics do you expect to change?