INFO 3501/5501: Lab Assignment #3

Due on Wednesday, October 26, 2016 at 9:00am

Professor Brian Keegan
Department of Information Science
College of Media, Communication, and Information
University of Colorado, Boulder

Background & Objectives

Relationships within Wikipedia data can also occur between different kinds of nodes. Lab Assignment #3 will explore different approaches to identify and quantitatively analyze the structure of users' contributions to articles as bipartite graphs.

1. **Combining data retrieved via API and MySQL**. This lab will continue to demonstrate the capabilities of using the combination of MySQL and API queries to identify and analyze relational data about Wikipedia. Students will extend functions from previous labs to retrieve and parse the content of Wikipedia pages.
2. **Analyzing the structure of bipartite networks.** As distinct from hyperlink network, a collaboration network is a unique kind of *bipartite* graph. Students will compute and interpret metrics specific to bipartite graphs such as clustering and assortativity.
3. **Visualizing the structure of bipartite networks**. `Gephi` can also be used to visualize the structure of bipartite graphs. Students will visualize the collaboration network of an article and the pages it links to as well as the editors who contribute to them.
4. **Interpreting projections of bipartite networks**. Bipartite graphs can be "projected" from two-mode into one-mode networks to use more traditional analytical methods. Students will learn how to interpret one-mode projections of the bipartite collaboration graphs by visualizing their backbones as well as comparing their structure to the structure of a hyperlink network.

Instructions

We will be be using Jupyter Notebooks running on the Wikimedia Foundation's PAWS infrastructure to retrieve public data from replica MySQL databases containing the complete history of revisions to articles as well as data exposed by the the Wikipedia API. We will use the account you created in Lab 1 as well as the `networkx` and `igraph` libraries and the `Gephi` program you installed in Lab 2.

Problem 1

Summarize how the structure of bipartite graphs (also known as "two mode" or "affiliation" networks) like a collaboration network is different from the hyperlink networks we examined in Lab #2. What are the two sets of nodes and what does a link represent? The lab notebook uses the hyperlink network to identify a larger set of articles for us to analyze. What does it mean for articles that are hyperlinked together to also share contributors? Discuss this in the context of other processes (motivation, identity, norms, *etc.*) we reviewed in the first five weeks of class. What are other potential relationships within Wikipedia besides hyperlinks we could use to identify a larger set of articles related to a single article? What kinds of data would we need to collect to be able to build out these non-hyperlink relationships? What would the collaboration graph of these articles linked by common editors potentially reveal about how Wikipedia works that is different from a hyperlink perspective?

Problem 2

Run all the code in the notebook until the "Compute descriptive statistics for the bipartite graph" section starts. This should have created several files: (1) "collaboration_<page_title>.gexf" is the bipartite collaboration graph of the hyperlinked articles and their editors, (2) "collab_page_bb_01_community_<page_title>.gexf" is the projected collaboration graph's backbone at the 0.01 threshold with community detection labels, and (3) "collab_page_bb_10_community_<page_title>.gexf" is the projected collaboration graph's backbone at the 0.10 threshold with community detection labels (remember that the "backbone" only retains the strongest edges for each node in the network). Load the bipartite collaboration GEXF file ("collaboration_<page_title>.gexf") up into Gephi and visualize your graph, making sure to color the node types (user vs. page) by color. Include one PNG file that showcases the structure of the network best based on different graph layout algorithms (ForceAtlas 2, Fruchterman-Rheingold, Yifan Hu, *etc.*) and sizing nodes by different attributes. Discuss some interesting features about this visualization in terms of what users or pages are particularly central, how some pages and their editors appear to cluster together, how interpretable the graph is overall, *etc.*

Problem 3

Run the notebook sections starting at "Compute descriptive statistics for the bipartite graph." Discuss the findings from the descriptive statistics for the bipartite network. Who are the most connected editors and what are the most connected articles in the network? What is surprising or to be expected among the most-edited articles? Speculate about the consequences of filtering out the bot users on these lists of most-connected users and the connectivity of the collaboration network as a whole. Do pages' or editors' connectivity show stronger evidence of a power law relationship? Why can we not measure the reciprocity in the collaboration graph like we did for the hyperlink graph?

Problem 4

Discuss the findings about the clustering and assortativity in the collaboration network. How should we interpret the clustering coefficient for a page in a bipartite graph like the collaboration network? Which pages had the highest 'dot'-clustering coefficients and why are these either to be expected or surprising? Is there a convincing relationship between the connectivity of a page and its clustering coefficient within the collaboration network? Include the scatterplot of the article degree vs. article clustering. Discuss what editor behaviors could be contributing to the observed relationship between articles with more editors having lower or higher clustering than articles with fewer editors. How should we interpret the average neighbor degree statistic calculated on pages versus users? Are there convincing relationships between the connectivity of a page or node and their average neighbor connectivity? Include the user and page degree vs. neighbor degree scatterplots and interpret the results.

Problem 5

Run the notebook sections starting at "Compute descriptive statistics for the projected graph". The collaboration network describing the user-page relationships was projected into a graph containing only pages. What are the relationships connecting the pages in the projected graph? How does the density and distribution of clustering values for the projected graph compare to the densities of your hyperlink network in Lab # 2 or the bipartite network above? Are the relationships between node degree and node clustering or avg neighbor degree different for the projected graph compared to the bipartite graph? Include images for both scatterplots and interpret their results.

Problem 6

We saw in the previous problem that the projected collaboration graph is significantly more dense and clustered than the bipartite or hyperlink networks we have visualized previously. We will use a statistical method developed by Serrano, Boguña, & Vespignani (2008) we will extract the "backbone" of the projected collaboration network containing the most "important" links for each node. Specifically, we will extract two "backbone networks" containing the top 10% and top 1% most important links and compare them visually. Load each of the "collab_page_bb_01_community_<page_title>.gexf" and "collab_page_bb_10_community_<page_title>.gexf" GEXF files you generated in previous steps into Gephi. Visualize them as you have done previously by employing a combination of graph layouts, node sizes by degree, and node colors by community membership. Include PNG files for both network layouts. How do these two backbone network visualizations compare to each other in terms of interpretability? Is the sparser top 1% network easier to understand or has it lost too many nodes to tell a complete story? Are these projection networks easier or more difficult to understand than the bipartite collaboration network you visualized in Problem 2? If you used the same Wikipedia article in Lab #3 as you did in Lab #2, how does the projected collaboration network compare to the hyperlink network in terms of community structure and prominent nodes?

Acknowledgements

I want to thank the Wikimedia Foundation for the PAWS system and related Wikitech infrastructure that this workbook runs within. Yuvi Panda, Aaron Halfaker, Jonathan Morgan, and Dario Taraborelli have all provided crucial support and feedback.

Licensing