Week 2 Part 2 - Assignment Centrality Measures

by Matthew Lucich

PROMPT

Centrality measures can be used to predict (positive or negative) outcomes for a node. Your task in this week's assignment is to identify an interesting set of network data that is available on the web (either through web scraping or web APIs) that could be used for analyzing and comparing centrality measures across nodes. As an additional constraint, there should be at least one categorical variable available for each node (such as "Male" or "Female"; "Republican", "Democrat," or "Undecided", etc.)

In addition to identifying your data source, you should create a high level plan that describes how you would load the data for analysis, and describe a hypothetical outcome that could be predicted from comparing degree centrality across categorical groups.

RESPONSE

The GoodReads dataset collection scraped by UC San Diego in 2017 is an expansive compilation of data, which includes various publicly available information from GoodReads.com. The data collection includes CSVs and JSONs files of books, users' shelves, and reviews. One of the datasets of particular interest, and one I have analyzed for Data 607, is the user-generated book genres dataset. I use the term "user-generated" and UCSD researchers use the term "fuzzy" to describe the genre dataset, since GoodReads does not assign genres to books but instead lists the genres users have associated the books with. After intensive data manipulation, the format that precedes the edge list can be previewed below (additional genres not pictured):

title	book_id	economics	politics	sociology	society	political-science	government	history
Eiger Dreams	10849	0	0	0	0	0	0	1
Great Masters: Mozart: His Life and Music (Gre...	3523380	0	0	0	0	0	0	1
The Earth Is Weeping: The Epic Story of the In...	31678077	0	0	0	0	0	0	1

When loading into a NetworkX graph object, book_id would become the nodes and would be assigned categorical variables indicating what genres they are associated with. The edges between the nodes would be assigned a weight that is the percentage of overlapping genres. For example, if ten genres are considered and two books share six out of the ten genres, then the edge weight would be 0.6. If the nodes (books) share no genres then no edge between them would exist.

An interesting analysis would be to see if certain books act as a bridge across genres. For instance, perhaps we could see behavioral economics books such as Thinking Fast, Slow and Nudge act as connectors between economics and psychology. Therefore, we would want to measure betweenness centrality. Following the same idea of searching for books that connect genres, we can also compute closeness centrality to see if particular books are highly close to other nodes due to their position between clusters. Though there can be various reasons for nodes (books) receiving high closeness scores.

References

Irfan, Mohammad (2020), Programming with NetworkX in Python, https://www.youtube.com/watch?v=CPQeSmDGiOQ

Mengting Wan, Julian McAuley (2017), Goodreads Datasets, https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home

Systems Innovation (2015), Network Centrality, https://www.youtube.com/watch?v=NgUj8DEH5Tc

Tsvetovat, Maksim & Kouznetsov, Alexander (2011), *Social Network Analysis for Startups*