Group Project
Math 168 Essay 4

Dhruv Chakraborty, Daniel Luo,
Pratyusha Majumder, Lisa(Yixuan) Shen

November 15, 2020

1 Abstract and Outline

The flexibility of networks to model phenomena ranging from citation networks to grocery store orders allows for uncovering several fascinating perspectives which may otherwise be overlooked. We are interested in the usage of Social Network Analysis in recommender systems, particularly methods using centralities and clustering metrics. Ronald Burt's notion of Structural Holes provides us with a starting point to consider "gaps" in networks that should be filled, in some sense (Ronald Burt's paper). Burt quantified his notion through the concept of Redundancy, which is closely related to the local clustering coefficient. We want to approach recommender systems through the lens of Structural Holes and extend the concept to include measures of centrality.

One key challenge in this will be identifying how the addition of an edge (or in some cases even a node) impacts a network. By quantifying improvements to a network through a constructed metric, we want to look at various manners in which we recommend additions to a network, and see which methods succeed with different types of networks. In particular, we believe Betweenness centrality could be interesting to look at, and we want to extend the notion of k-cores to a proper centrality metric which we think will also be a good way to recommend changes. The interpretability of such changes is also of specific interest to us. We want to look at a broad range of networks - some of which include citation networks (could a paper have benefitted from citing another paper?), an Instacart grocery orders dataset (is there another

item the shopper may have wanted to buy?) and a Caltech Facebook dataset (which we will prototype everything on).

A Brief Outline is listed below:

1. Read current literature and research
    - Solidify our theoretical understanding of how k-core centrality and structural holes are calculated and in what use cases they're applicable
    - Broadly research other types of centrality measures to see which ones are applicable in addition to core centrality
2. Brainstorm how to create a metric of our own
    - Discuss how we can mathematically create some sort of metric that encapsulates centrality with respect to k-core centrality and structural holes
3. Apply newly created metric to actual data
    - Break down how to calculate the new metric we created given some network leveraging Python, its packages or whatever is easiest (potentially MATLAB)
    - Apply the algorithm to calculate metrics to Caltech Facebook dataset
    - See what insights it gleans and how it compares to other types of centrality measures
    - Apply algorithm to other datasets listed in part 4 and see what insights it gleans
4. Work on final report and presentation
    - Create a write up summarizing the insights
    - Create visualizations to help audience gain intuition on what we found
    - Create slide deck to present our findings

2 Plans For Individual Responsibilities

As a group, we have decided to work on the project in close collaboration with each other so that everyone is able to learn and contribute to the project. In terms of network data analysis, the tentative distribution of tasks for data is as follows: Dhruv will work on data cleaning, preprocessing, and sampling; for data exploration and visualization, all of us will work on the prototype Caltech Facebook data set together and each of us individually will explore/visualize our own data set of interests. For composition of the final paper, as a group we will conduct literature searches together with each person focusing on different topics. Daniel will write the introduction, Lisa and Pratyusha will be responsible for writing and editing the main content sections, which Daniel and Dhruv will also assist on especially with the analysis of their own networks of interests, and Dhruv will write the conclusions. All of us will attend the group presentation and Daniel has volunteered to make the slides for the presentation.

3 What We Hope to Get Out of the Project

Daniel - I'm hoping to learn more technical skills in terms of manipulating and handling network data practically whether it's in Python or some other programming language or software. This is so I have the skills in the future to perform my own network analysis. I'm also intrigued by the idea of creating a new "concept" in math with creating a different centrality metric.

Dhruv - I'm looking forward to learning more about the subtle differences between centrality metrics and any particular features they may highlight. I'm also really excited to work more closely through code with networks, to create a k-core clustering metric and to compare the value of any recommender systems we may come up with.

Lisa - I want to expand my knowledge on the relationship between different centrality measures and redundancy and see how that can be applied to an analysis of social influence within the context of a recommender network. I also would like to gain more technical skills of network data analysis (i.e. data cleaning/processing/visualization).

Pratyusha - I am hoping to develop my technical skills in Python, as it can be applied to network analysis. I am also hoping to further learn conduct exploratory analysis and expand my understanding of centrality measures

and structural holes in the context of data that can potentially be used to build and improve recommender systems.

4 Data Sets that We Plan to Use

One dataset we want to work on is the Caltech Facebook Dataset that consists of 709 nodes representing users on the social network and edges representing friendships between those users [4]. We chose this dataset for our initial investigation because the notion of Structural Holes has extensive applications in social network analysis and using such a social network could give us interesting insights about the phenomenon. More specifically, identifying structural holes in the network might give us insight into which profiles to recommend to certain users that may have a potential to be "friends". These findings could then help expand the network of Facebook friends so advertisers can access new users that were previously unknown to them.

After our initial investigation, we want to see how we can analyze other types of networks using clustering and centrality measures. Applying these measures on public transportation networks will help us identify transit sites that have the highest degrees and the ones that appear the most when identifying a shortest path between two locations. These findings could potentially help travellers better plan their journeys by determining transit locations that are the most important to have access to. Some of the public transportation networks we are interested in working with are the ones provided by the Kujala et. al [2].

The centrality measures and concept of structural holes will also be interesting to investigate on citation networks. Using DBLP+ Citation and the ACM Citation Network dataset would help us identify the papers that are the most commonly cited and papers and pairs of papers where one does not cite the other, yet they both cite other highly connected papers [5]. This could be the result of one paper being published before another but will also give rise to questions such as "Will one paper have benefited from citing the other paper, given that they both cite a small cluster of highly connected papers?" We're also interested in working with Instacart data to use centralities with structural holes to create a recommendation system specifically for Instacart data [1]. Most recommendation systems in this domain will recommend singular items or some sort of ranked list of items given some data about what items are in the cart. This can be seen through something

like Facebook's friend recommendation algorithm. However, we're hoping that applying this concept of structural holes will give way to recommending many items at a single time given some items in the cart. This would be better suited for use-cases where someone isn't potentially a single user but perhaps a larger entity (e.g. corporations and larger organizations) that can buy many items at once so that a recommendation system that bulk recommends items is better than a recommendation system that recommends individual items.

The Amazon co-purchasing network is composed of 262111 nodes representing products and 1234877 directed edges depicting which products were co-purchased frequently [3]. Amazon's product recommendation system recommends product j to a user who views product i, for a directed edge from node i to j. Using our centrality measures with the structural holes will help us identify the products that are co-purchased with the greatest number of products and products that may not be commonly purchased with other products but have high connectivity to a common cluster. This can be especially useful to improve the recommendation for new products that are introduced on the platform.

The Movielens dataset is one also worth exploring, using the centrality techniques mentioned above. This bipartite network is composed of nodes representing users and movies, and a weighted edge connecting them, with the edge weights representing the rating the user has assigned to the movies[4]. A standard recommender system would potentially use movie tags and user ratings to determine which movies are recommended to a particular user. However, after conducting a one-mode projection of this bipartite network to obtain groups of users that have enjoyed the same movie; we can quantify enjoyment as a metric based on the edge weights in the original bipartite network. Then, the various centrality measures and the concept of structural holes can be used to determine the users that have the most common interest in movies with other users (users found with high k-core centrality). The concept of structural holes can be used to predict which users may potentially have a similar taste in movies that has yet to be established. Streaming services can then leverage this idea to modify their recommender systems to propose movies to users that may not have previously been chosen by their standard system.

References

[1] The instacart online grocery shopping dataset 2017. Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on November 15, 2020.

[2] Kujala, R., W. C. D. R. e. a. A collection of public transport network data sets for 25 cities. *Sci Data*, 5 (2018).

[3] Leskovec, J., Adamic, L. A., and Huberman, B. A. The dynamics of viral marketing. *ACM Transactions on the Web 1*, 1 (May 2007), 5.

[4] Rossi, R. A., and Ahmed, N. K. The network data repository with interactive graph analytics and visualization. In *AAAI* (2015).

[5] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. Arnetminer: Extraction and mining of academic social networks. In *KDD'08* (2008), pp. 990–998.