Final Project Report

Daneiel Yalew, Matthew Orr, Phillip Tezaur

2024-05-29

1 Introduction

The Epinions dataset is a who-trust-whom online social network of a general consumer review site Epinions.com. Members of the site can decide whether to "trust" each other. The network consists of 75879 nodes and 508837 edges.

Analyzing this network can provide insight to the role of trust in online forums. The measurement of key metrics, such as the average degree of a given node, or the mean distance across the network, can provide insight on how users of other online forums may interact with each other. Furthermore, the visualization of this network, its communities, most central nodes, and adjacency matrix can allow us to more clearly identify trends in the network.

2 Methodology

The data processing involved was relatively straight forward. The edge-list for the network was downloaded from the SNAP dataset repository, and read into a DataFrame object in R. Using the `igraph` package, the data was transformed to an `igraph` object. The `igraph` package extends base R functions such as `plot` to allow for the interaction with the `igraph` object. Additionally, the package provides functions that allow for the measurement of metrics such as edge density, connectivity, etc., which will be investigated further below (§3).

3 Analysis

3.1 Induced Sub-Graphs

Four induced subgraphs, consisting of a random sample of 200 nodes each, are visualized below (Fig. 1). Only a small number of connections exist in each subgraph, indicating that there is weak connectivity between the nodes in each graph.

3.2 Network Metrics

The network has been summarized by various metrics, as displayed in the table below.

Mean Distance	Edge Density	Reciprocity	Transitivity	Diameter	Connected Components?	Articulation Points	Modularity
4.754723	8.83774e-05	0.405226	0.06567883	16	No, 42176 nodes unconnected	15936 nodes	0.39

Each metric can be described as follows:

- Density: The ratio of edges in a graph to the total number of edges in the graph.
- Reciprocity: The ratio of total number of reciprocal edges relative to the total number of edges.
- Transitivity: The tendency of nodes to cluster.
- Reachability: The aggregate version of node-level reachability; the ability of one node to reach other nodes in a directed graph.
- Diameter: The longest possible shortest path between any pair of nodes in the graph.
- Connected components: The subgraphs within the graph such that all nodes are within the subgraph are connected by a path. Separated into strongly connected components and weakly connected components.

- Articulation points: The nodes that if removed would split the graph into different component sub-graphs.
- Modularity: The measure of the density of connections within a communities; the communities which are *greedily* formed with the `cluster_fast_greedy()` algorithm, included with the `igraph` package.

3.3 Community Detection

As stated above, the `cluster_fast_greedy()` algorithm *greedily* forms communities within the network and attempts to find subsequent nodes which fit into these "communities". In the Trust Network, over 3200 communities were found. The 5 largest communities had 24051, 18925, 9590, 902, and 898 nodes, respectively. This tells us that the 5 largest groups contain just under 75% of all of the nodes in the network.

3.4 Adjacency Matricies

The initial adjacency matrix (Fig. 2) is very sparse, this is due to the fact that there are a significant amount of nodes, each with a relatively small degree.

TODO: Include adjacency matrix for reordered nodes.

4 Results

- Key findings from the analysis
- Visual and textual summaries

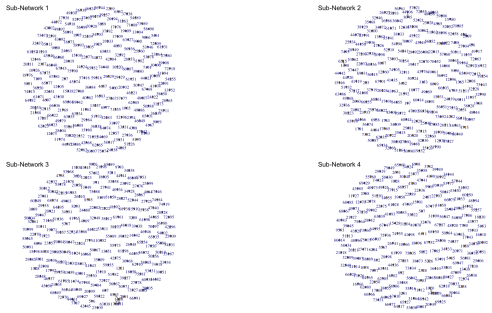

Figure 1: Induced Sub-Graphs of the Trust Network.

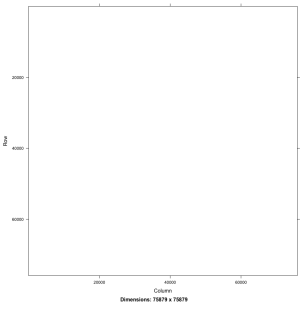

Figure 2: Initial Adjacency Matrix.

5 Conclusion

- Summary of insights gained
- Possible future directions for further analysis

6 References

https://snap.stanford.edu/data/soc-Epinions1.html

7 Appendix

7.1 Code