Assortative Mixing in the Amazon.com Book Network

Ben Collingsworth and Ronaldo Menezes
BioComplex Laboratory
Computer Sciences, Florida Tech, Melbourne, Florida, USA
bcolling@my.fit.edu, rmenezes@cs.fit.edu

Introduction

Online book reviews are important because, depending on review content, they may influence whether consumers will purchase a book. Of further interest, the set of books a person reads and reviews reveals information about that person. People select a book to read with the expectation that they will enjoy it. This expectation is based on their background, affiliations, and biases. Accordingly, the set of book reviewers may be viewed as homogenous since groups of people share common elements of background, affiliation, and biases. For example, a community of book reviewers may be found in people that share a common passion for wine. They read and reviewbooks(23,54),(897,995) describing wine biographies, and tasting. This work explores the presence of homogeneity/heterogeneity in online book reviews as reflected in a complex network constructed from books and book reviews from Amazon.com. In complex network analysis, homogeneity is characterized by the property known as assortativity which measures the "feather" phenomenon where people tend to associate with other people that are similar to themselves. Understanding of assortativity within book reviewer networks provides sociological insight: are people balanced in the materials they read or do they tend stay within a narrow range of their inclinations? In addition, knowledge in this area has important marketing applications including improved marketing strategies and prediction of the books people read.

Related Work

Assortativity coefficient is a measure of assortative mixing for networks containing vertices that are classified by attribute. Newman defines an assortativity coefficient that can be calculated for any network. He begins by defining basic proportional values that describe how the ends of edges are distributed among types of vertices:

1. \( r_{ij} = 1 \): The fraction of all edges that join vertices of type \( i \) with all vertices of type \( j \).
2. \( r_{ii} = a_i \): The fraction of edges that join a vertex of type \( i \) with a vertex of type \( i \).
3. \( r_{ij} = b_{ij} \): The fraction of edges that join a vertex of type \( j \) with a vertex of type \( i \).

Using these fractions, the level of assortative mixing in a network is quantified by the equation:

\[
r = \frac{\sum_i c_i r_{ii} - \sum_i \sum_j c_i c_j r_{ij}}{\sum_i c_i (1 - \sum_j c_j r_{ij})}
\]

The assortativity coefficient has a range of \(-1 \leq r \leq 1\). A value of 1 indicates perfect assortativity where all edges are of the same type. A value of 0 indicates no assortative mixing. A negative value for \( r \) tells us that the network is disassortative.

Experimental Methods

Data for the book reviewer network was collected using a Web crawler on the Amazon.com website. The Web crawler runs recursively to collect books and reviewers. Each level of recursion starts with a list of books. The collection algorithm starts with a set of "seed" books. Information on each book is collected and stored in a database including the set of reviewers that reviewed each book. The list of books for the next level of recursion is built from the set of "other" books that add a reviewer also reviewed. For the analysis performed, the algorithm was limited to three levels of recursion because of the exponential growth of the data set. Beyond these levels, the data set becomes too large to effectively analyze.

The Book Network

The book network was constructed from the two bipartite tables in the database: 1) books, and 2) reviewers (see Figure). In the network, individual books are represented as vertices. If two books are reviewed by the same reviewer, the books are connected by an edge. Edges are undirected and weighted by the number of times the pair of books is listed as a book that was "also bought" for the other book in the vertex pair. To see the correlation, the percentage of edges in the "Also Bought" Network that were also in the Book Network was taken. Again, the Book Network was filtered at various levels by edge weight to test the strength of the correlation. The results show that there is a commonality of ~90 percent when no filtering is done. This result is a validation of the Book Network because it shows that connections in the Book Network correspond to the "Also Bought" Network generated by an external algorithm that links books through common readers. Further, the results show a commonality of 40 percent even when the Book Network is highly pruned. This is an indication that connected books in the "Also Bought" network are indeed strongly related because they have many connections in the Book Network.

Network Assortativity

Assortativity in the Book Network was examined using various levels of filtering. The network was filtered by edge weight to highlight important vertices and edges. Edges not having the desired minimum weight were removed from the network. Additionally, vertices that were left with a degree of zero were removed. Four levels of filtering were applied for the analysis with edge weights \( \geq 5 \), \( \geq 10 \), \( \geq 20 \), and \( \geq 40 \). Also, the original network with no filtering was included. The book category assigned by Amazon is used as the attribute to measure assortativity. Thirty categories were found during network creation. The figure below shows that with no filtering the network is slightly disassortative. However, as less important vertices and edges are removed, the network is shown to become much more assortative. With a filtering level of 40, the network is highly assortative with an assortativity coefficient of 0.65. The disassortativity seen with low levels of filtering may be attributed to noise produced from many connections between books that few reviewers have read and reviewed.

Conclusions

The Book Network exhibits classic social network properties with a small average path length, high average clustering coefficient, and high degree probability coefficient of 1.9. Further, the overall network structure was validated through correlation with the "Also Bought" Network which was generated directly from Amazon.com data. Preliminary results indicate the strong presence of assortativity in the Book Network. This assortativity is prominent when the network is filtered to remove unimportant vertices and edges. Further research to be conducted includes:

- Correlation of network measurements sales rank
- Analysis of the data from a different viewpoint with reviewers represented as vertices, and edges between the reviewers that have read the same books.
- Profiling of reviewers: do reviewers who read one type book also tend to read another type of book?