David Moste, Euclid Zhang, Samuel Reeves

Week 2 Part 2 - Assignment Centrality Measures

The Data

The data set we have found for this assignment is the Amazon product co-purchasing network metadata network from Stanford. The group variable for each node is a categorical variable (Book, DVD, Video, or Music).

The data is already stored in a dictionary format, which will help with loading. Once brought in, each product (node) has a unique number associated with it (ASIN) that can be used as an identifier. There is also a group category, ASINs of co-purchased products, and a list of reviews containing customer ID, date, and rating. The full data format along with an example can be seen below:

Data format:

- **Id**: Product id (number 0, ..., 548551)
- **ASIN**: Amazon Standard Identification Number
- **title**: Name/title of the product
- **group**: Product group (Book, DVD, Video or Music)
- **salesrank**: Amazon Salesrank
- **similar**: ASINs of co-purchased products (people who buy X also buy Y)
- **categories**: Location in product category hierarchy to which the product belongs (separated by |, category id in [])
- **reviews**: Product review information: time, user id, rating, total number of votes on the review, total number of helpfulness votes (how many people found the review to be helpful)

```
Id:   15
ASIN: 1559362022
  title: Wake Up and Smell the Coffee
  group: Book
  salesrank: 518927
  similar: 5  1559360968  1559361247  1559360828  1559361018  0743214552
  categories: 3
   |Books[283155]|Subjects[1000]|Literature & Fiction[17]|Drama[2159]|United States[2160]
   |Books[283155]|Subjects[1000]|Arts & Photography[1]|Performing
Arts[521000]|Theater[2154]|General[2218]
   |Books[283155]|Subjects[1000]|Literature & Fiction[17]|Authors, A-Z[70021]|( B
)[70023]|Bogosian, Eric[70116]
  reviews: total: 8  downloaded: 8  avg rating: 4
    2002-5-13  cutomer: A2IGOA66Y6O8TQ  rating: 5  votes:   3  helpful:   2
    2002-6-17  cutomer: A2OIN4AUH84KNE  rating: 5  votes:   2  helpful:   1
    2003-1-2  cutomer: A2HN382JNT1CIU  rating: 1  votes:   6  helpful:   1
    2003-6-7  cutomer: A2FDJ79LDU4O18  rating: 4  votes:   1  helpful:   1
    2003-6-27  cutomer: A39QMV9ZKRJXO5  rating: 4  votes:   1  helpful:   1
    2004-2-17  cutomer:  AUUVMSTQ1TXDI  rating: 1  votes:   2  helpful:   0
    2004-2-24  cutomer: A2C5K0QTLL9UAT  rating: 5  votes:   2  helpful:   2
    2004-10-13  cutomer:  A5XYF0Z3UH4HB  rating: 5  votes:   1  helpful:   1
```

The Process

The ASINs of co-purchased products are already in a useable format to become edges!

Ideally, the customer IDs of reviewers would be turned into nodes as well. To do this, we would need to run through the reviews portion of each product and extract the customer ID to check for existing nodes

with that ID. If none existed, we would create a new node. Each customer would have an edge for all products they reviewed and a category labeled customer to distinguish them from products.

Prediction

A possible outcome that could be predicted using degree centrality would be to check if products or customers have more ties within a network. Products are obviously related to each other in many ways and have a guaranteed base of other similar products. However, unlike products, customers have the ability to seek out other customers and extend their network on their own. This could provide information on whether customers purchasing results are more influenced by other customers or product recommendations.