# Amazon Referral Network

- Rohit Hiwale
- Sanjeev Chandra Vadde
- Satish Chandra

Mentor - Joydeep Chandra

# Problem Statement

- Collect data for books from www.amazon.com
  - For each book, data includes :
    - Title
    - Authors
    - Ratings
    - List of books bought together
- Analyze the properties of network of books derived from data
  - Network has
    - Nodes : Books
    - Edges : Relations such as
      - same author
      - bought together

# Questions to be answered

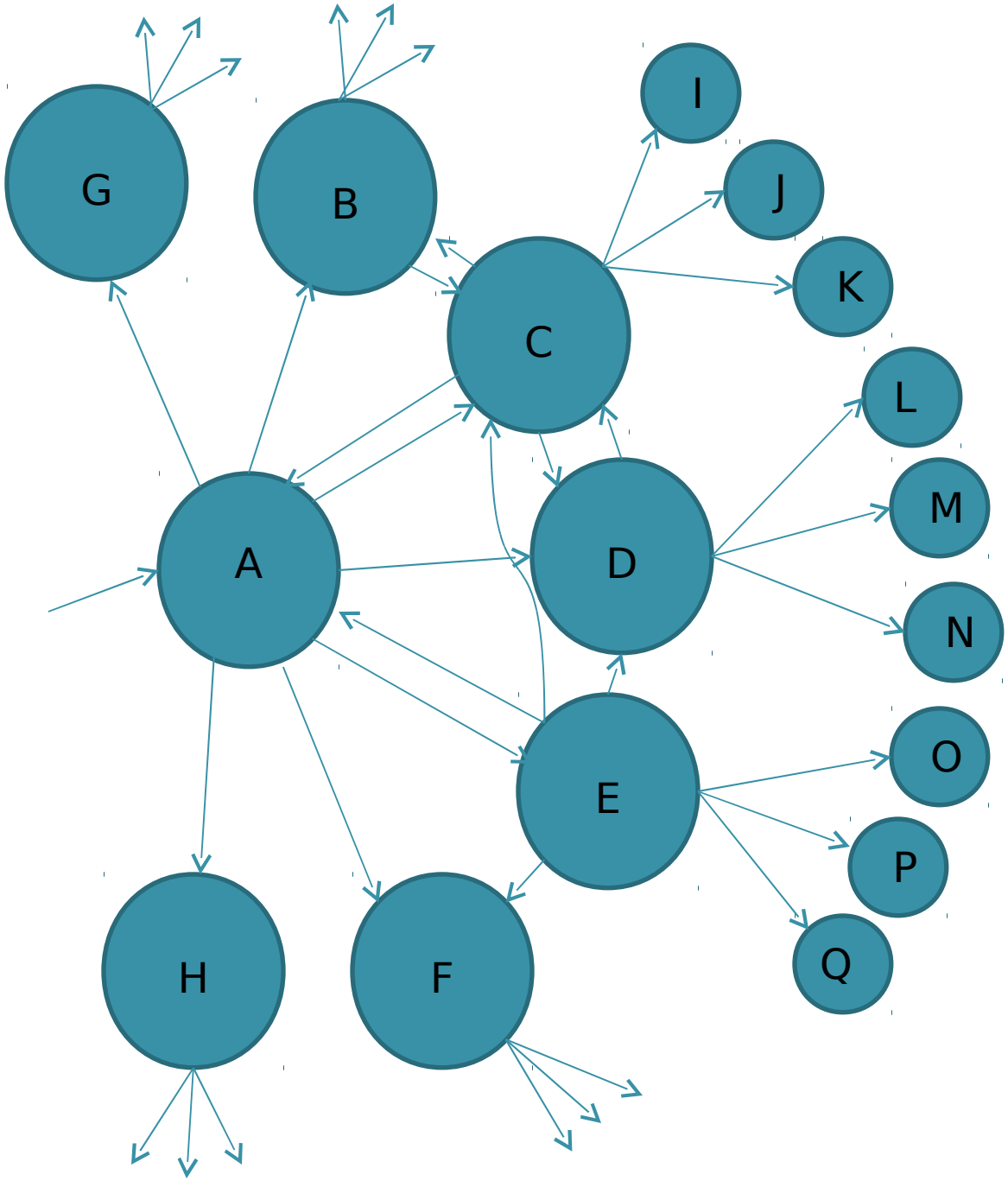
- Will a book likely to be popular if any one of its author is popular?
- Books co-authored by popular authors are likely to be more popular than books written by a single popular author?
- For co-author network and co-customer network, discuss:
  - Degree Distribution
  - Clustering Co-efficient
  - Assortativity

# Network Topology

- Each book recommends 7 other books bought by the same customers.
- Each other books recommend a set of 7 books themselves. The process repeats.
- Multiple instances of the same book being encountered.
- Book already exists – Discard
- Book doesn’t exist in the graph - Add

# Network Topology(contd..)

- Edges between books with a common feature.
- In this case, feature refers to
  - a same author (or)
  - a recommendation of one book on another's page.
- An example looks like,


G
B
I
J
K
C
L
M
A
D
N
E
O
P
H
F
Q

# Will a book likely to be popular if any one of its author is popular?

- r = #(popular books with popular authors) / #(popular books)

| Rating Threshold for popularity (out of 5) | Ratio (r) |
|---|---|
| 3.5 | 0.992 |
| 4.0 | 0.948 |
| 4.5 | 0.870 |
| 4.7 | 0.784 |
| 4.8 | 0.768 |

**From the above data, it can be concluded that as there are many books which are highly popular and yet do not have an already popular author.**

# Books co-authored by popular authors are likely to be more popular than books written by a single popular author?

- r1 = #(popular books by multiple popular authors) /#(books by multiple popular authors)
- r2 = #(popular books by single popular author) /#(books by single popular author)

| Rating Threshold for popularity | Multiple Popular Authors (r1) | Single Popular Author (r2) | Difference \|(r1-r2)\| |
|---|---|---|---|
| 4.0 | 0.968 | 0.881 | 0.087 |
| 4.5 | 0.935 | 0.801 | 0.134 |
| 4.7 | 0.957 | 0.893 | 0.064 |
| 4.8 | 0.902 | 0.920 | 0.018 |

- An interesting observation
  - The difference between the two ratios rises a bit and falls.
  - From the observation it can be concluded that
    - Books with less ratings need multiple authors to get popular.
    - Books with very high ratings have no such constraint.
    - Hence, for books of high ratings, it doesn’t matter that they have a single author or multiple authors.

# Analysis of the book popularity given a single popular author vs. many less-popular authors

- r1 = #(popular books by single popular author) /#(books by single popular author)
- r2 = #(popular books by multiple less-popular author) / #(books by multiple less-popular author)

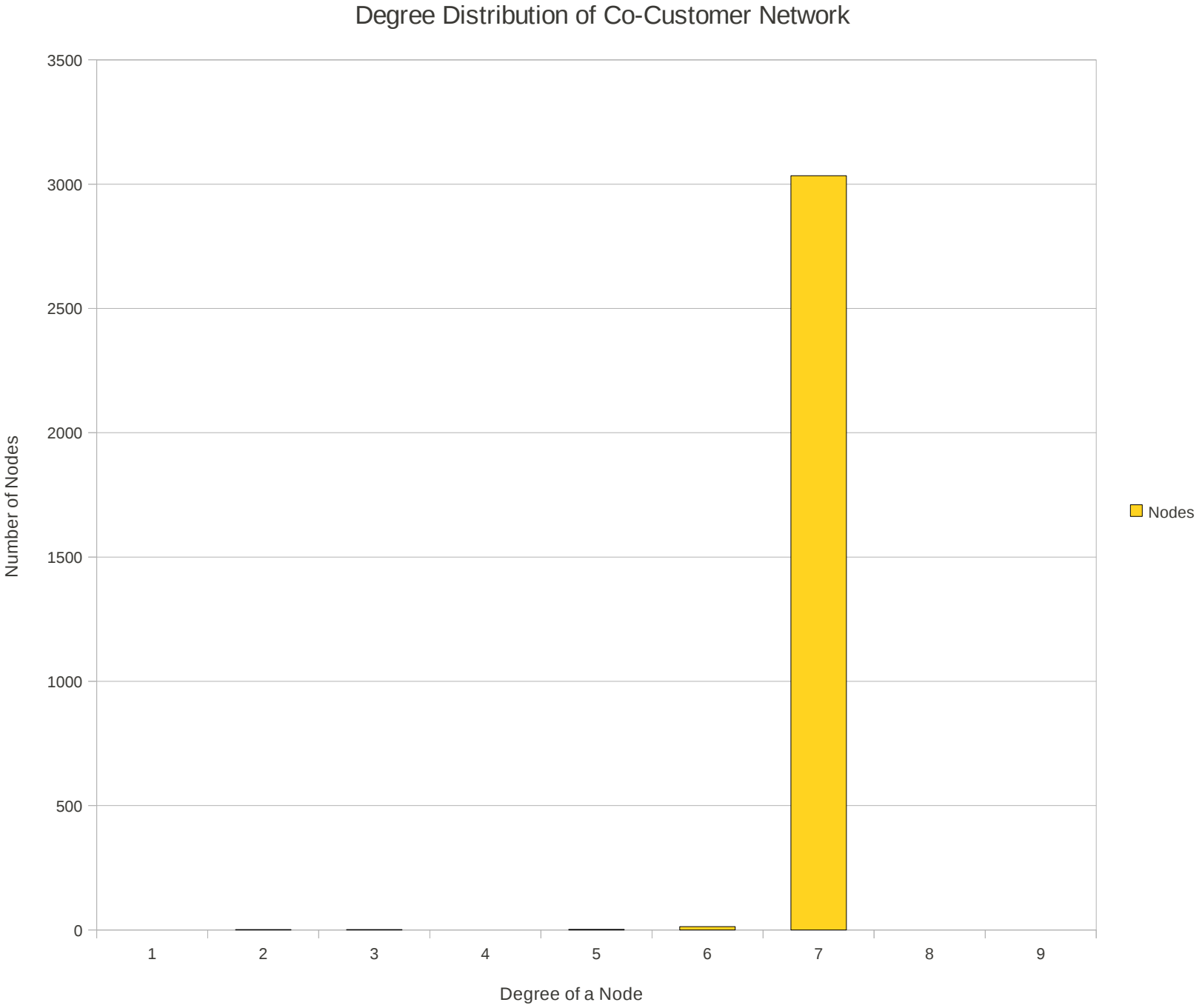
| Rating Threshold for popularity | Single Popular Author (r1) | Multiple Less-Popular Authors (r2) |
|---|---|---|
| 4.0 | 0.881 | 0.062 |
| 4.5 | 0.801 | 0.025 |
| 4.7 | 0.893 | 0.041 |
| 4.8 | 0.920 | 0.029 |

From the data, it is evident that,

- Single popular author makes a book more popular than many less-popular co-authors.

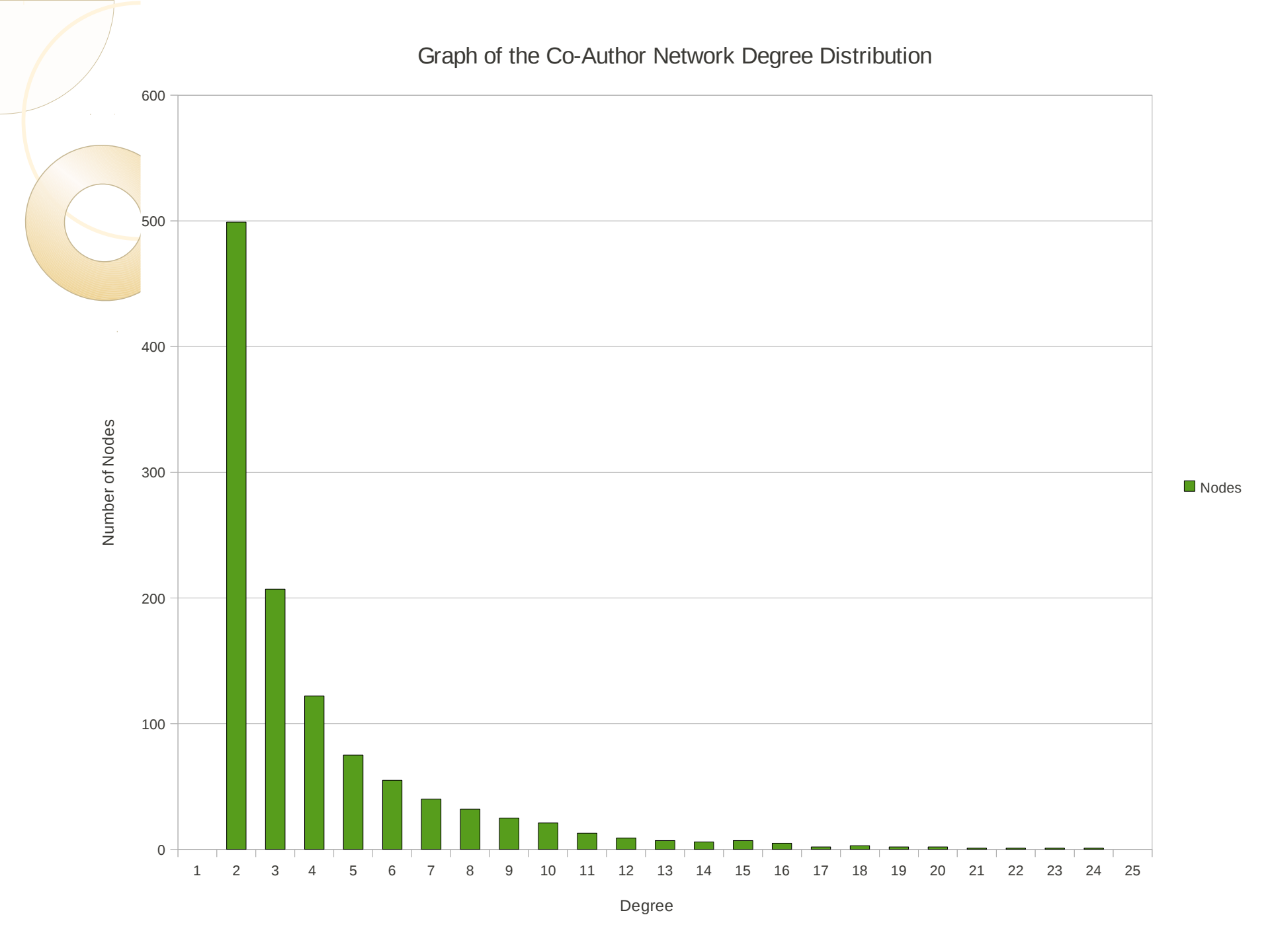
# Degree Distribution

- Co-customer network :
  - Amazon gives 7 “also recommended books” for each book entry.
  - Thus, a degree of 7 for more than 99% of the nodes.
  - Interesting fact :
    - **Amazon prevents the existence of a 3-edged loop spanning over 3 layers in this graph.**
  - This may be a part of Amazon's strategy for some suitable content discovery.
  - Thus, contributing to the growing network, but hiding the underlying actual network.


Degree Distribution of Co-Customer Network
Number of Nodes
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
Degree of a Node
Nodes

# Degree Distribution

- Co-author network :
  - Different from the co-customer network.
  - The degree falls gradually from a large value and goes into oblivion.
  - Inference:
    - Number of books with less co-authored books is hugely more than the number of books with more co-authored books.


Graph of the Co-Author Network Degree Distribution
Number of Nodes
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Degree
Nodes

# Clustering Co-efficient:

- Co-author network :
  - Observed to be around 0.779.
  - Many nodes have individual CC values of 1.
  - Inference:
    - It can be inferred that the graph is a set of cliques connected by edges.
    - The nodes with less value of CC are the nodes which form connections between two cliques.

# Clustering Co-efficient (Contd.)

- Co-customer network :
  - Neighborhood of each node is almost same due to typical topology of the graph.
  - CC for each node is approx 0.714
  - Hence, CC for whole graph is also approx 0.714

# Assortativity

- Co-author network :
  - Observed assortativity = 0.785394
  - Inference:
    - any book co-authored with a high degree book is likely to have high degree too.
    - similarly, any book co-authored with low degree books is likely to have a low degree.

# E[i][j] : fraction of edges between nodes of degree i and j

```
>>>
>>> print_Eij( E )
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.06175 0.00128 0.00270 0.00028 0.00043 0.00014 0.00014 0.00000 0.00000 0.00028 0.00000 0.00000 0.00028 0.00000 0.00000 0.00000 0.00000 0.00043 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00128 0.08167 0.00085 0.00341 0.00043 0.00071 0.00000 0.00000 0.00000 0.00014 0.00028 0.00028 0.00000 0.00000 0.00000 0.00000 0.00000 0.00028 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00270 0.00085 0.06858 0.00028 0.00384 0.00000 0.00028 0.00000 0.00057 0.00000 0.00000 0.00057 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00028 0.00341 0.00028 0.04297 0.00057 0.00413 0.00000 0.00000 0.00028 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00043 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00043 0.00043 0.00384 0.00057 0.04013 0.00228 0.00228 0.00000 0.00000 0.00000 0.00057 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00014 0.00071 0.00000 0.00413 0.00228 0.04952 0.00142 0.00085 0.00142 0.00014 0.00100 0.00000 0.00000 0.00142 0.00000 0.00000 0.00000 0.00014 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00014 0.00000 0.00028 0.00000 0.00228 0.00142 0.04582 0.00000 0.00398 0.00185 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00085 0.00000 0.05635 0.00000 0.00541 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00228 0.00000
0.00000 0.00000 0.00000 0.00057 0.00028 0.00000 0.00142 0.00398 0.00000 0.03415 0.00413 0.00697 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00100 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00028 0.00014 0.00000 0.00000 0.00000 0.00014 0.00185 0.00541 0.00413 0.05891 0.00185 0.00256 0.00000 0.00000 0.00000 0.00000 0.00000 0.00014 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00028 0.00000 0.00000 0.00057 0.00100 0.00000 0.00000 0.00697 0.00185 0.01423 0.00341 0.00128 0.00000 0.00000 0.00000 0.00000 0.00014 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00028 0.00057 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00256 0.00341 0.02134 0.00199 0.00740 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00028 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00128 0.00199 0.00057 0.00142 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00142 0.00000 0.00000 0.00000 0.00000 0.00000 0.00740 0.00142 0.06346 0.00000 0.01053 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00341 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.02220 0.00000 0.00555 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.01053 0.00000 0.00285 0.00000 0.00000 0.00000 0.00000 0.00000 0.00028 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00555 0.00000 0.00171 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00043 0.00028 0.00000 0.00043 0.00000 0.00014 0.00000 0.00000 0.00100 0.00014 0.00014 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.05407 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00228 0.00000 0.00000 0.00000 0.00000 0.00341 0.00000 0.00028 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00028 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.07854
>>>
```

# Assortativity (Contd..)

- Co-customer network :
  - Observed assortativity is approximately equal to 0.
  - The graph is a well connected graph with almost all nodes having the same neighborhood connectivity pattern.
  - Inference:
    - any book recommended by the users of a popular book need not be popular.

# Tools Used

- Data Collection
  - Python module mechanize to ignore robot.txt
  - Python module BeatifulSoup and pyparsing to parse html content
- sqlite3 database to store crawled data and as data structure for network.

# Problems Faced

- Amazon discourages crawling by
  - malforming html tags
  - introducing unrecognized Unicode characters
- Unstable internet
- Hence slow data collection.

Thank You.
Any Questions??