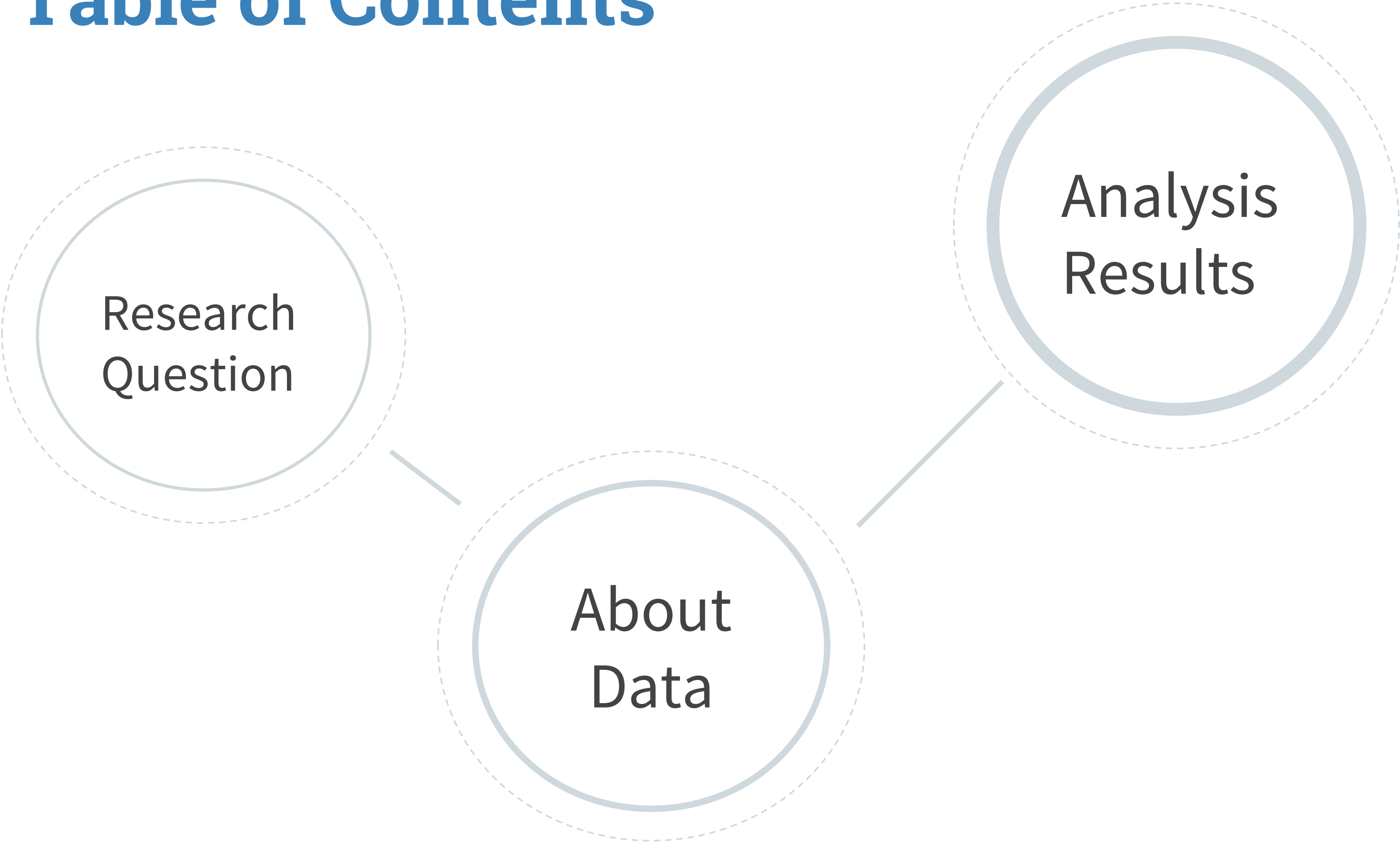
The YouTube Social Network

Agnes Liu, Lancy Mao, Robin Liu

Table of Contents



Research Question

About Data

Analysis Results

Research Questions

- Examine the contact network structure
- Examine the centrality measures of network and the correlations between them
- Examine whether the friend contact network follows the Power Law
- Examine correlations and relationships of attributes

Data Information

YouTube

Data background

- The original data set is retrieved from the ASU Social Computing Data Repository. (a nodes file and five edge files)
- The data set contains
  - Number of Nodes: **13723**
  - Number of edges: **76765**
  - Number of attributes: **5**
  - Missing Value: **NO**

Attribute information

1. The **contact network** between all users in the dataset
2. The number of **shared friends** between two users
3. The number of **shared subscription** between two users
4. The number of **shared subscribers** between two users
5. The number of **shared favorite videos** between two users

Data source: R. Zafarani and H. Liu, (2009). Social Computing Data Repository at ASU [http://socialcomputing.asu.edu]. Tempe, AZ: Arizona State University, School of Computing, Informatics and Decision Systems Engineering.

Analysis Method

- To examine user(vertex) properties, we calculated **centrality measurements**, including **degree, betweenness, closeness, and eigenvector** and analyzed their distribution and correlation.

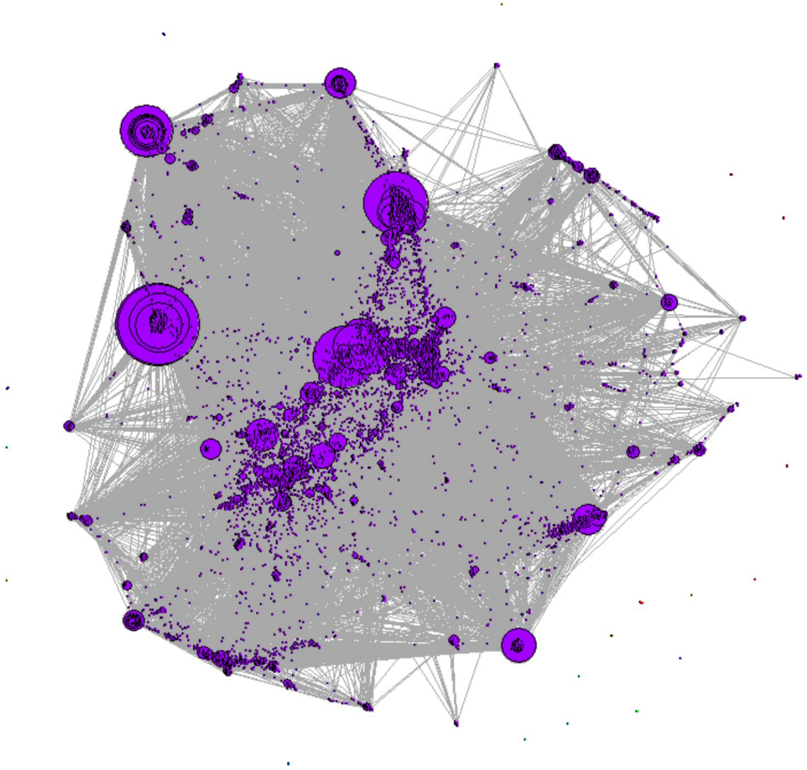
- To find out the network structure, we calculated **cluster** of network and **coreness** of every vertex. To measure network property, we also used network density and shortest path in the giant cluster.

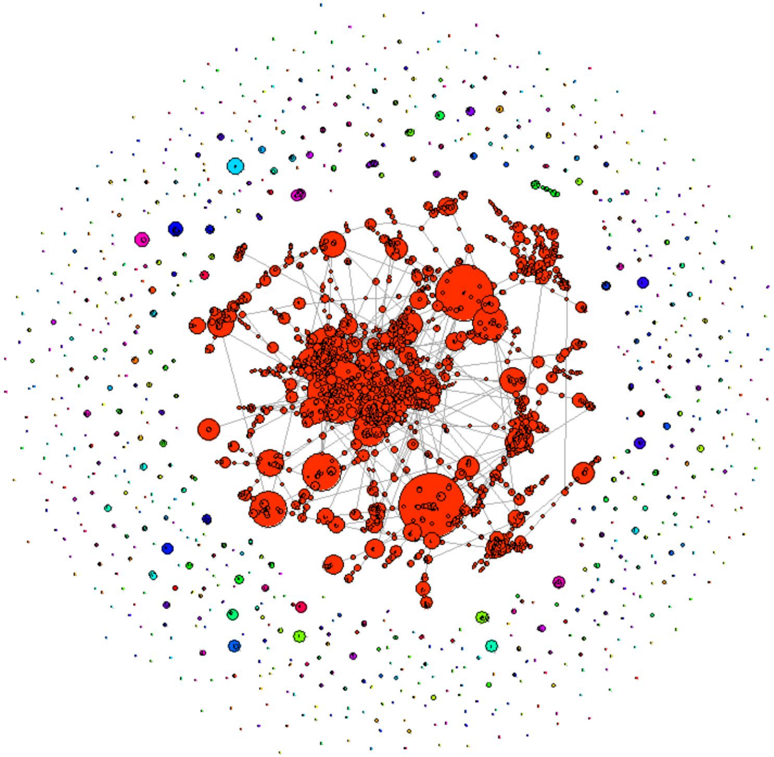
- **Attribute** distribution and correlation are used to measure attributes’ influence.

ANALYSIS RESULTS

YouTube Network

Network Graph

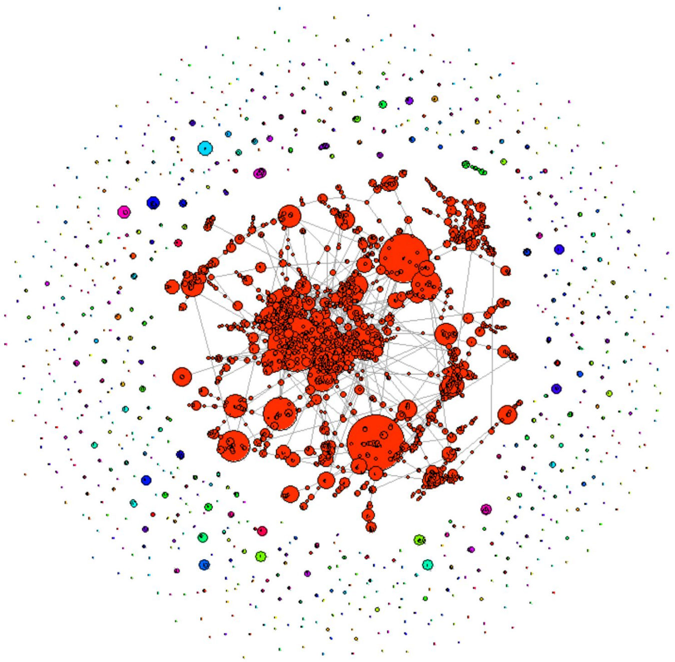


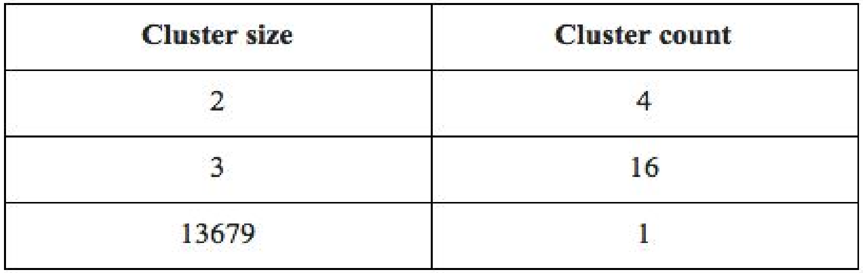


Network Sample Graph
(Sample=5000)

The Contact Network Analysis

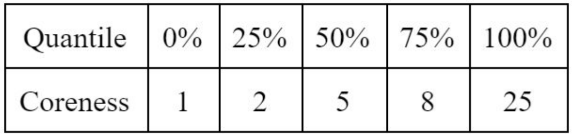
Examine the Network Structure





Cluster size	Cluster count
2	4
3	16
13679	1

- There are 21 clusters in total, including a giant one that contains over 99% nodes.
- From the result, we can tell that even though there is a giant component in the cluster, the nodes are connected to each other through multiple coordinators instead of a single gatekeeper.
- Distribution of coreness:



Quantile	0%	25%	50%	75%	100%
Coreness	1	2	5	8	25

Half of nodes could at least form a 5-core subgraph, which verifies our hypothesis that people can form connections with others who are not necessarily well connected, manage to reach to everyone and become a part of this giant component.

The Contact Network Analysis

Examine the Centrality Measures

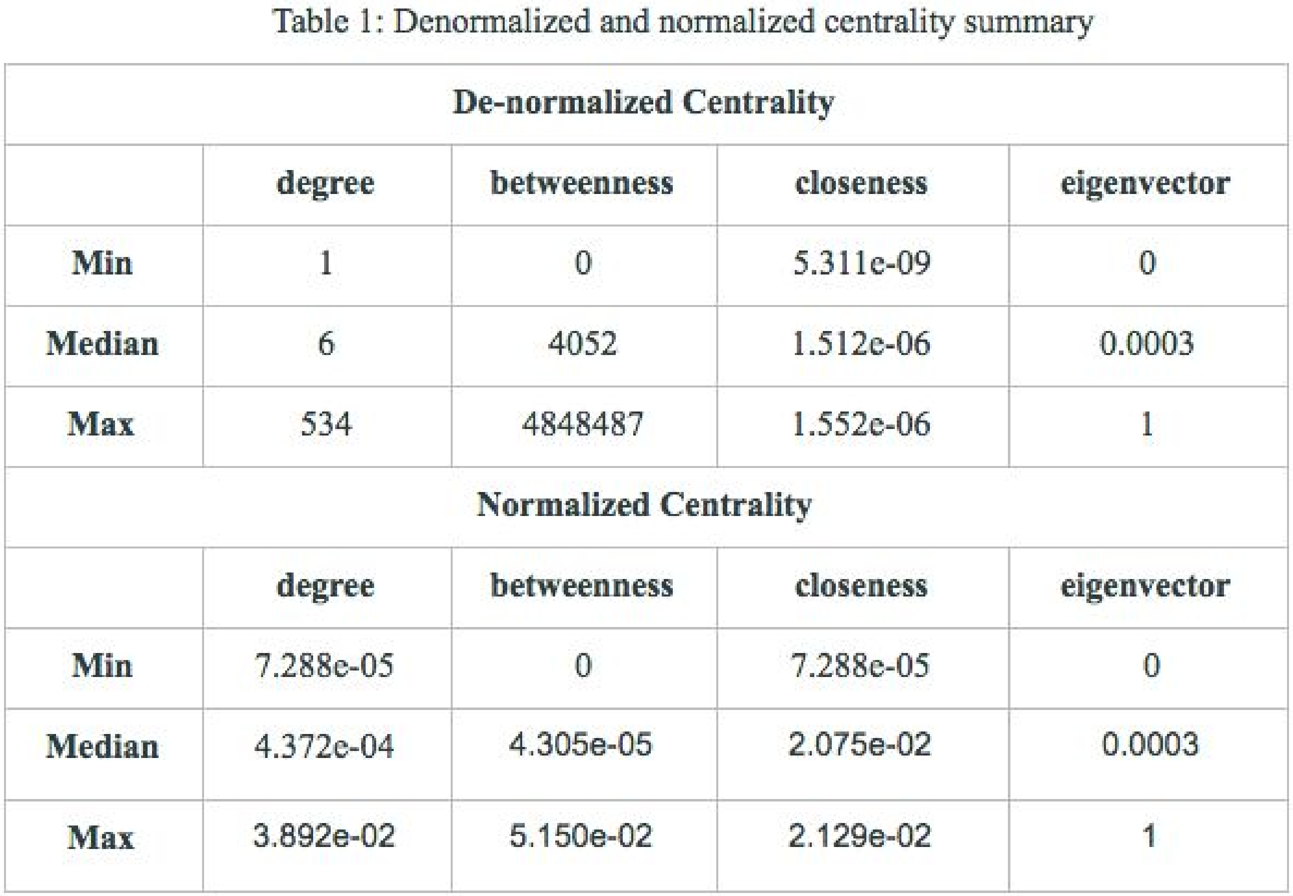


Table 1: Denormalized and normalized centrality summary

De-normalized Centrality				
	degree	**betweenness**	**closeness**	**eigenvector**
Min	1	0	5.311e-09	0
Median	6	4052	1.512e-06	0.0003
Max	534	4848487	1.552e-06	1
Normalized Centrality				
	degree	**betweenness**	**closeness**	**eigenvector**
Min	7.288e-05	0	7.288e-05	0
Median	4.372e-04	4.305e-05	2.075e-02	0.0003
Max	3.892e-02	5.150e-02	2.129e-02	1

These four centrality measurements show how central and important the users are

The Contact Network Analysis

Examine the Centrality Measures Correlation

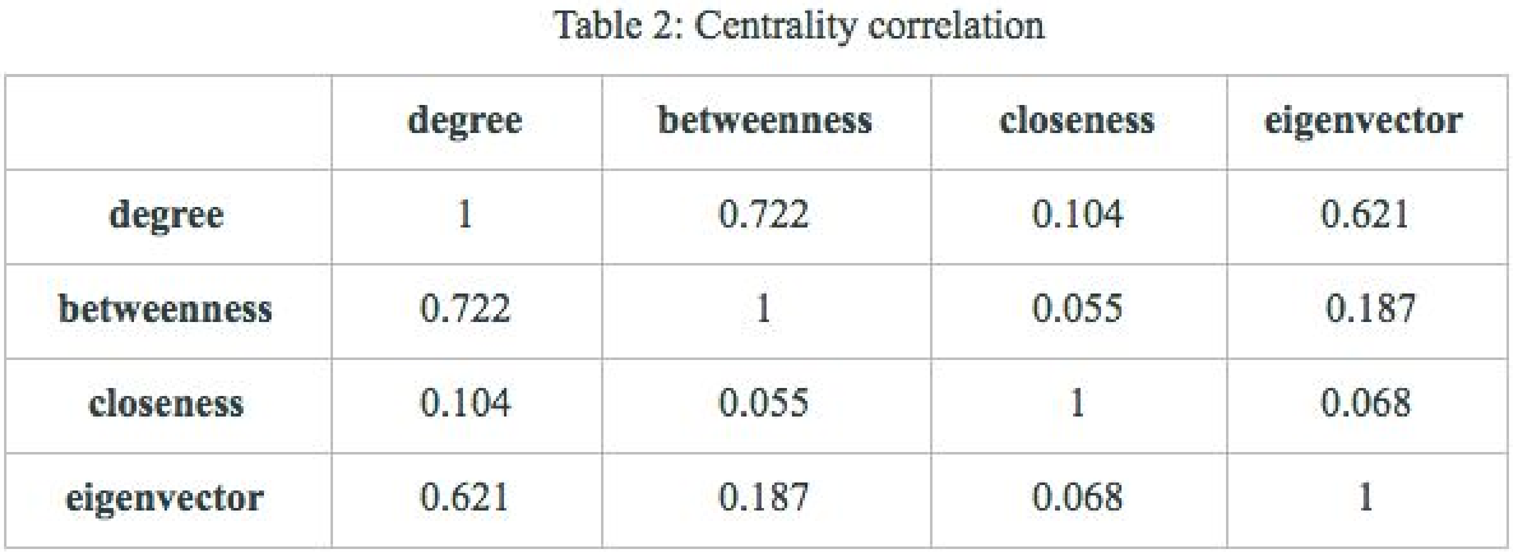


Table 2: Centrality correlation

	degree	betweenness	closeness	eigenvector
degree	1	0.722	0.104	0.621
betweenness	0.722	1	0.055	0.187
closeness	0.104	0.055	1	0.068
eigenvector	0.621	0.187	0.068	1

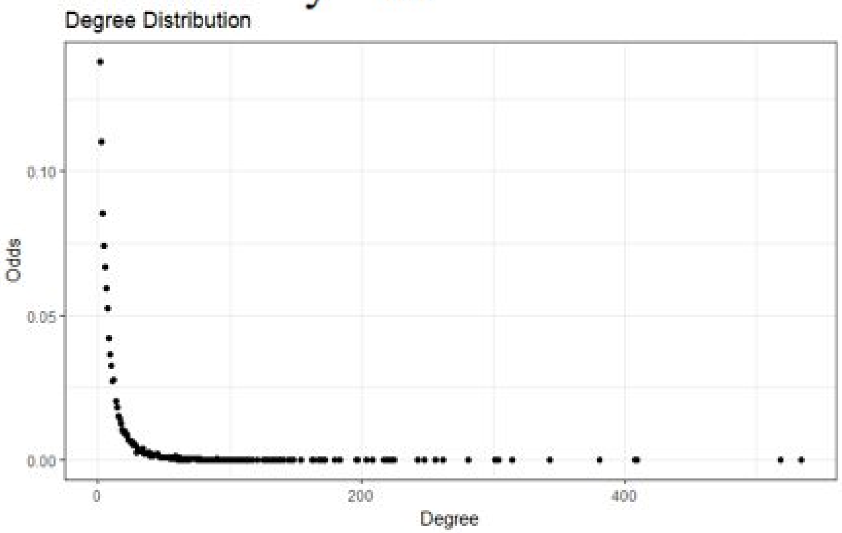
- Correlation between **degree** and **betweenness** (0.722) is the **highest.** High scores on betweenness centrality are associated with high scores on degree.
- Correlation between **degree** and **eigenvector** is 0.621. Users with more friends tend to connect with more central users.
- Correlations between **closeness** and **other three centrality** are very **low**. Users with high betweenness don't tend to situated near all other users, so they may act as gatekeepers among different clusters.

The Contact Network Analysis

*Examine whether the contact network follows **Power Law***

Our dataset reflects this phenomenon and follows the **Power Law**: Most of the people would have small degree centralities, only a handful of people would have large degree centralities.

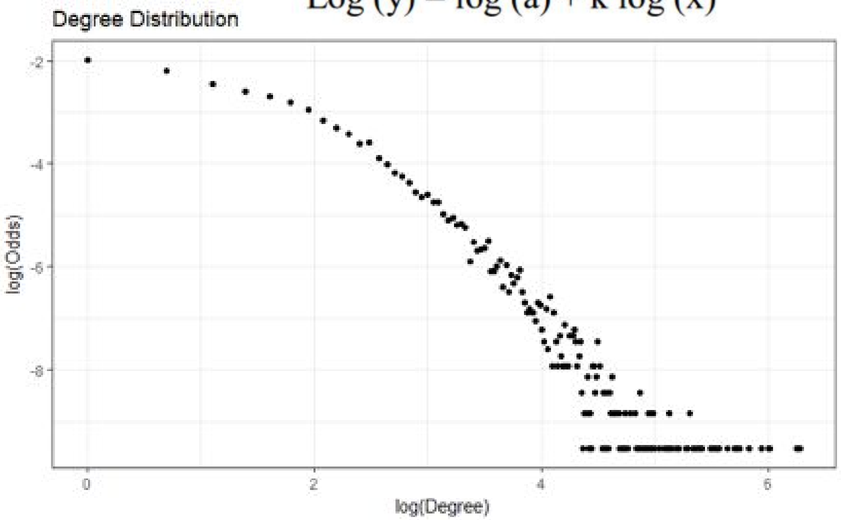
$$y = ax^k$$



The above graph shows that the **odds of a vertex** has some **inverse relationship** with the number of degrees.

$$\text{Log}(y) = \text{Log}(a\,x^k)$$

$$\text{Log}(y) = \log(a) + k\log(x)$$

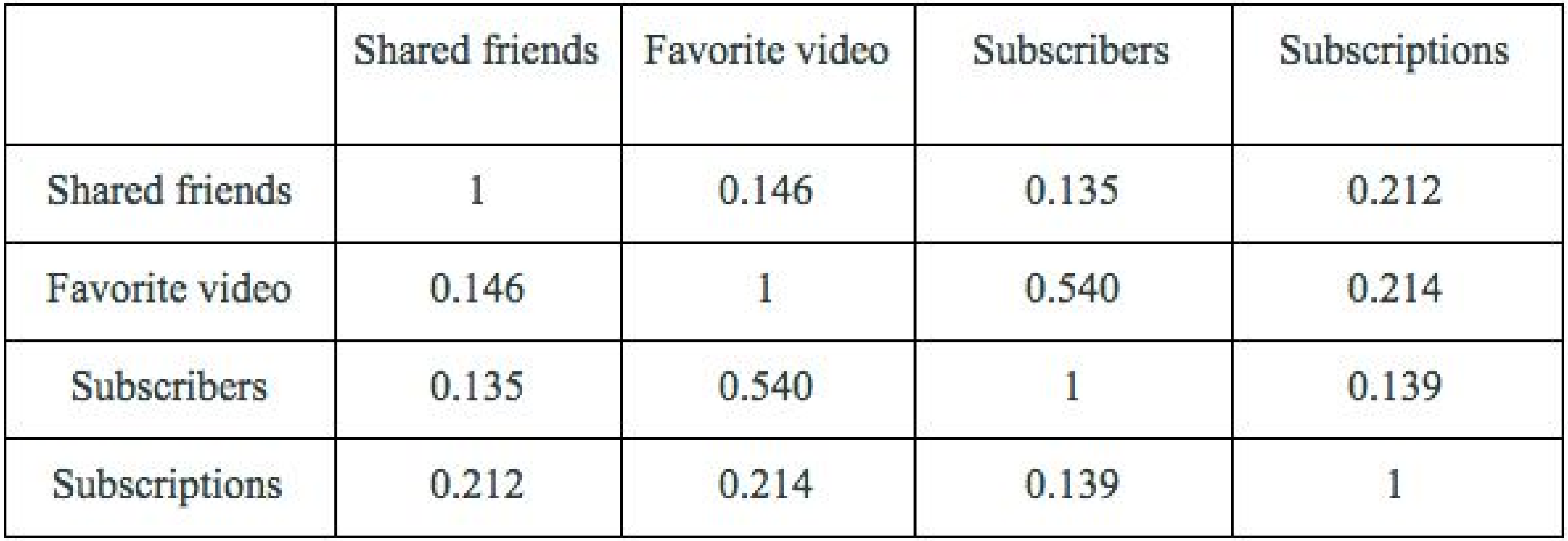


The above graph shows a linear relationship between the **log odds** and **log(degree)**, which indicates that the odds and number of degrees follow an **exponential decay relationship**, and our network follows the power rule.

Network Attributes Analysis

Examine the correlation between attributes

In general, correlations among these four attributes are low. Among all correlations, the correlation between **subscription** and **favorite videos** (0.54) is the highest. Generally, attributes don't vary together.



	Shared friends	Favorite video	Subscribers	Subscriptions
Shared friends	1	0.146	0.135	0.212
Favorite video	0.146	1	0.540	0.214
Subscribers	0.135	0.540	1	0.139
Subscriptions	0.212	0.214	0.139	1

Network Attributes Analysis

Examine the correlation between attributes

Question: Whether two users have shared friends/favorite videos/subscribers/subscriptions if they are connected?

We calculated **the number of edges** in the network which have attributes not equal to zero, indicating that the **shared attribute exists.** The following are our findings:

- 26.9% connections have shared subscribers.
- 46.8% connections have shared subscriptions.
- 46.5% connections have shared favorite videos.
- 16.1% connections have over 5 shared friends.

There are 50% probabilities that the user and his/her connected friend have shared subscriptions and favorite videos, while only 27% and 16% probabilities that they have shared subscribers and friends respectively.

Thank you!

Questions?