Sex Degrees of Separation: The Hollywood STD Network

MAS961 Networks and Complexity: Final Project
Otkrist Gupta, Julia Ma, Phil Salesses

The general population is fascinated by the lives of celebrities: where they go, what they do, and who they sleep with. So while the celebrity social network is well documented by tabloids and gossip columns, the actual network itself has not been modeled or studied. We are looking at various ways to visualize the celebrity sex network and derive interesting facts from keen observation. We will study the propagation of Sexually Transmitted Diseases through this network and calculate the probability of each celebrity having certain STDs.

Initially we attempted manual collection of data, but fortunately we found that many websites collect celebrity relationships. For example, Who's Dated Who (http://www.whosdatedwho.com/) lists thousands of celebrities, their dating and relationships profile, confirmed sexual encounters, and the respective time frames. We wrote a Python script that crawled this website, retrieved each page, and parsed them for relevant information. We seeded the crawler with Charlie Sheen and stored all of his sexual connections from his profile in our database. From there, the crawler systematically went through each connection to get more connections, cross-checking the database for duplicates, and gathered all links until the network was exhausted. This data was exported to Excel and then rendered this network using Cytoscape and Gephi. We were also able to automatically capture other relevant information, such as thumbnail profile images or birth dates, by modifying the web crawler.

Charlie Sheen is an actor best known for his role in the TV show "Two and a Half Men" and one of the highest paid TV actors of all time. We chose Charlie Sheen as our seed because at the time of this project, he was on the front page of the news for getting fired from "Two and a Half Men" for his alcoholism, drug abuse, and marital problems. He is also known for his multiple marriages and divorces and extremely high number of relationships (rumors say he has had 5000+ relationships), thus we thought he would be an excellent start of an expansive network of Hollywood. Using him as a seed, we gathered ~16000 nodes with ~21000 links.

Analysis of the Sex Network

Once we had the data, we used Cytoscape and Gephi to render and analyze the full network (Figures 1-8). Results from initial analysis are in Table 1. The graph is highly connected. Interestingly, by gathering data for the Hollywood sex network, we also retrieved a large portion of the porn star sex network. While we did not expect this, it is not surprising that these networks are interconnected since porn stars are a part of the movie industry. We also found the nodes that have the highest degree and those participating in the most number of 3-cliques, 4-cliques, and 5-cliques (Tables 2-5).

Number of nodes (using Charlie Sheen as the seed)	15991
Number of links	21482
Highest degree node	Rocco Siffredi with 114 links
Average path length	8.833
Average number of neighbors (average number of sex partners)	2.687
Number of partner swap motifs (4-cycle)	13733

Table 1. Results from initial analysis of full network.

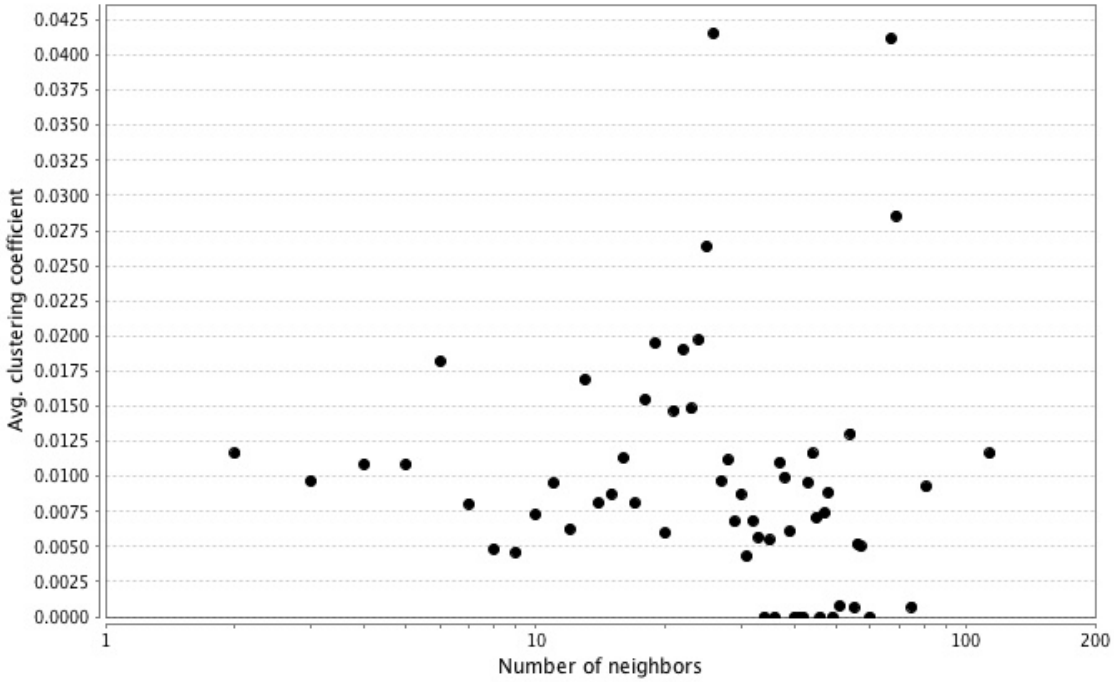


Figure 1. Graph showing the number of neighbors vs the average clustering coefficient. The higher the clustering coefficient, the more likely the nodes are to cluster together and become cliques. We expect the clustering coefficients to be higher than a random graph since people in the acting industry who have acted together in one movie are likely to act together in more movies. (This is just from observation. For example, movies with sequels tend to use the same actors and some movie directors frequently use the same actors, like Tim Burton with Johnny Depp.). In this graph, we see that there are a few nodes that, regardless of degree, are extremely clustering and some that are extremely non-clustering, while the majority average around 0.01 clustering coefficient.

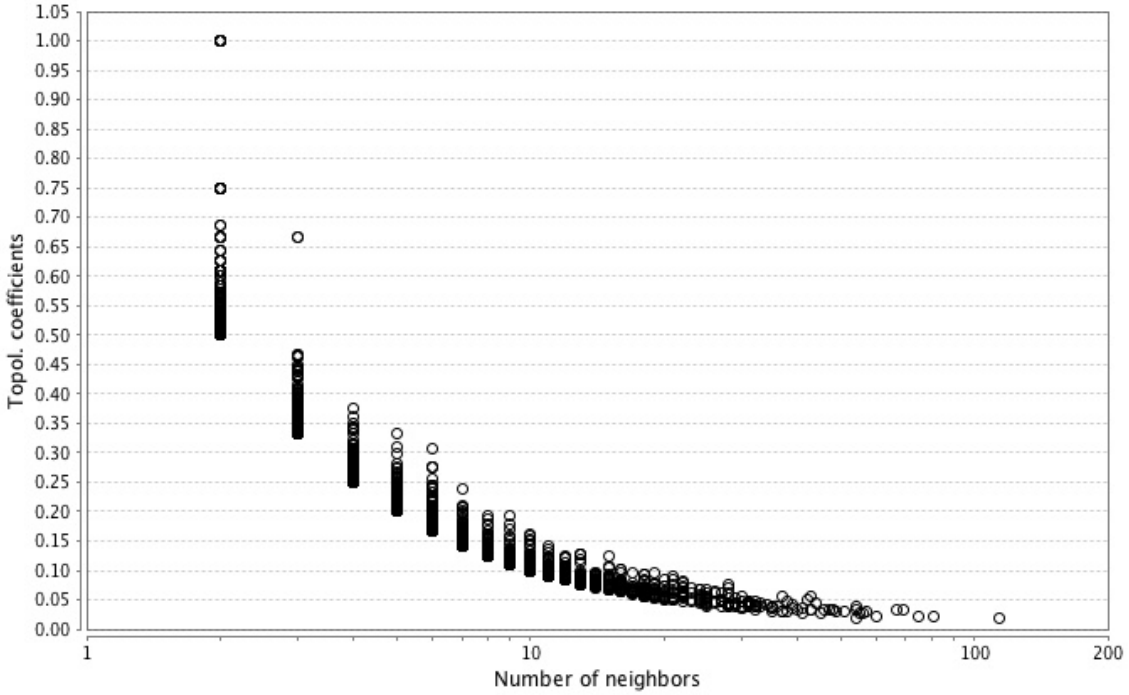


Figure 2. Graph showing number of neighbors vs topological coefficients. The topological coefficient gives a measure of whether nodes share neighbors or not. We see the trend that the higher the degree, the smaller the topological coefficient. We speculate this is because the higher degree nodes share a smaller percentage of its neighbors with nearby nodes.

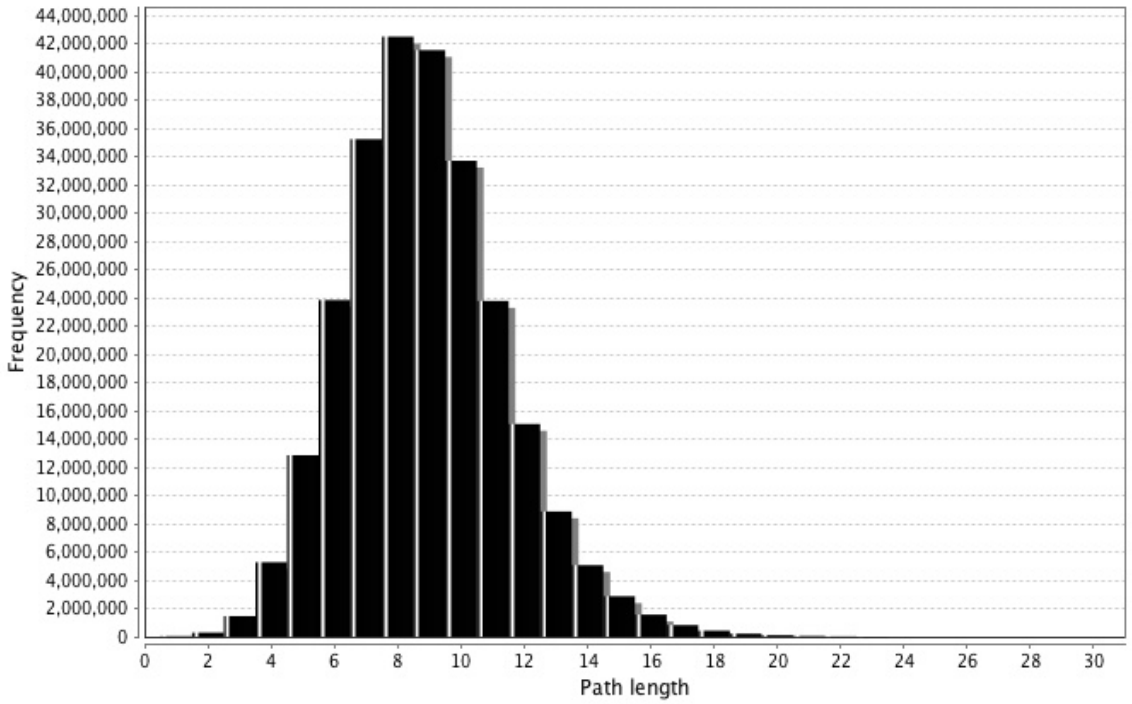


Figure 3. Frequency of N-length paths. Paths of length 8 and 9 are the most common, both occurring more than 40 million times. This distribution follows a Gaussian curve.

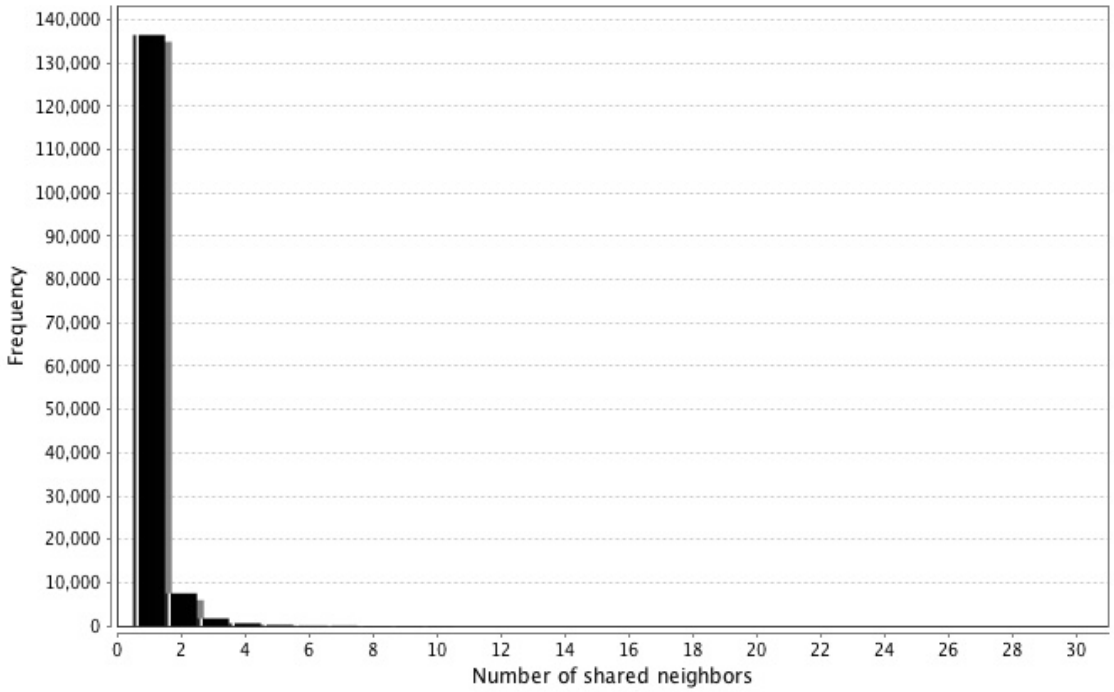


Figure 4. Frequency of the number of shared neighbors. This graph is highly skewed as having one shared neighbor is extremely common.

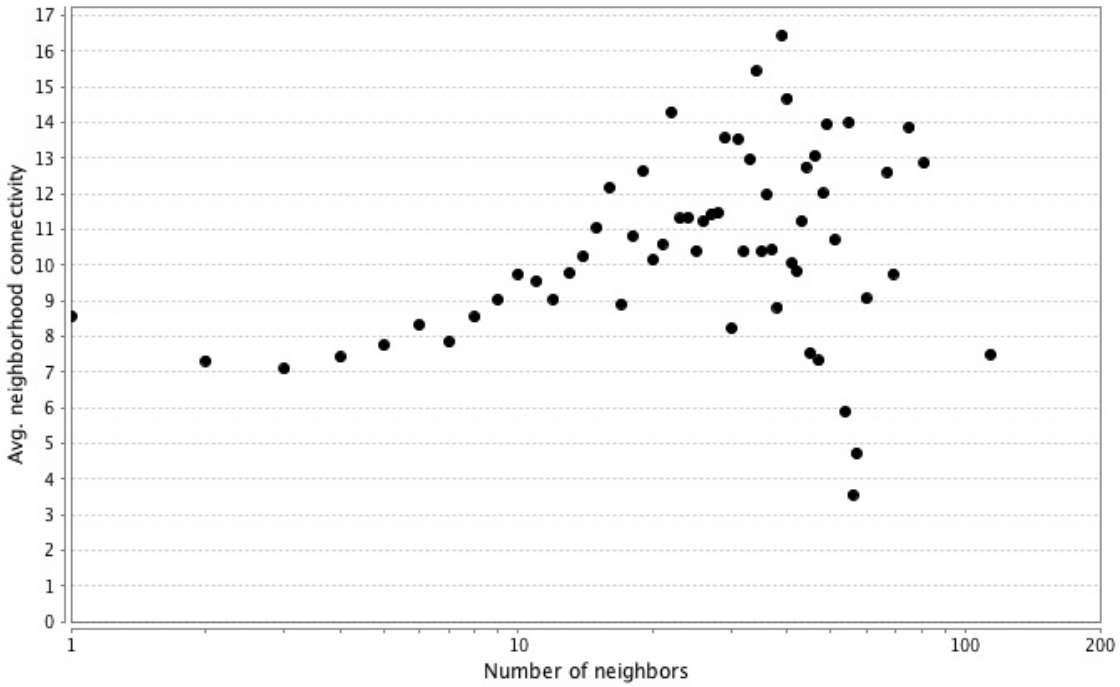


Figure 5. Number of neighbors vs average neighborhood connectivity. The neighborhood connectivity is defined as the average connectivity (degree) of a node's neighbors. We see that in general, the higher degree of a node, the higher degree of that node's neighbors.

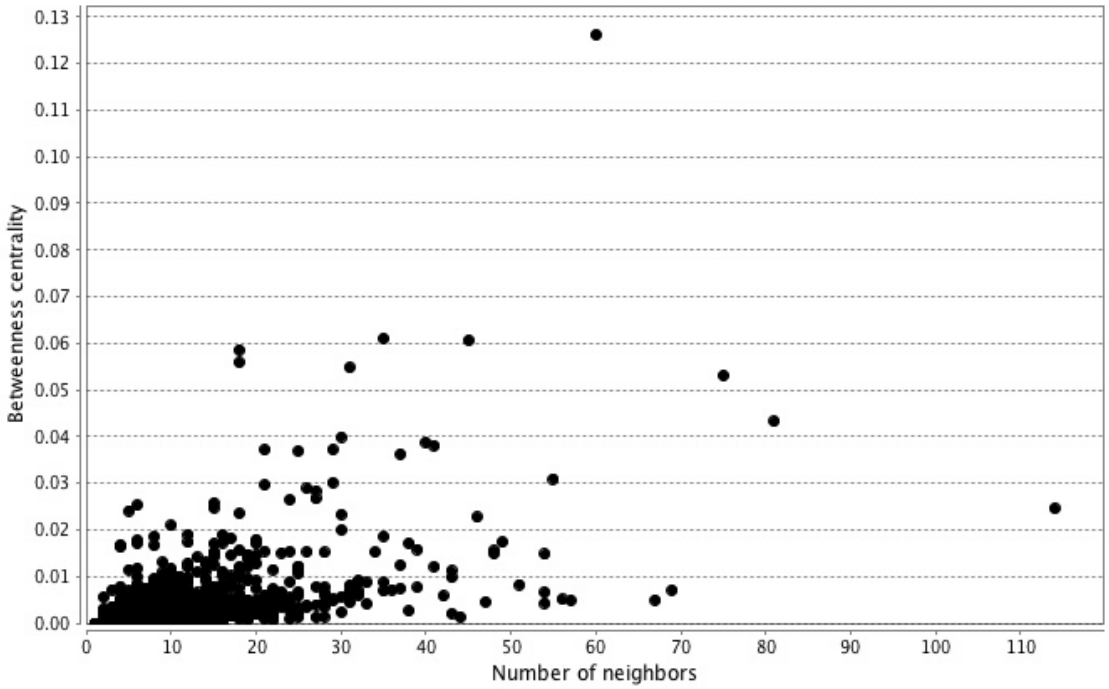


Figure 6. Number of neighbors vs betweenness centrality. Betweenness centrality counts the number of shortest paths that pass through a node. The majority of our nodes have very similar betweenness centralities, but in general, the higher the degree, the higher the betweenness centrality.

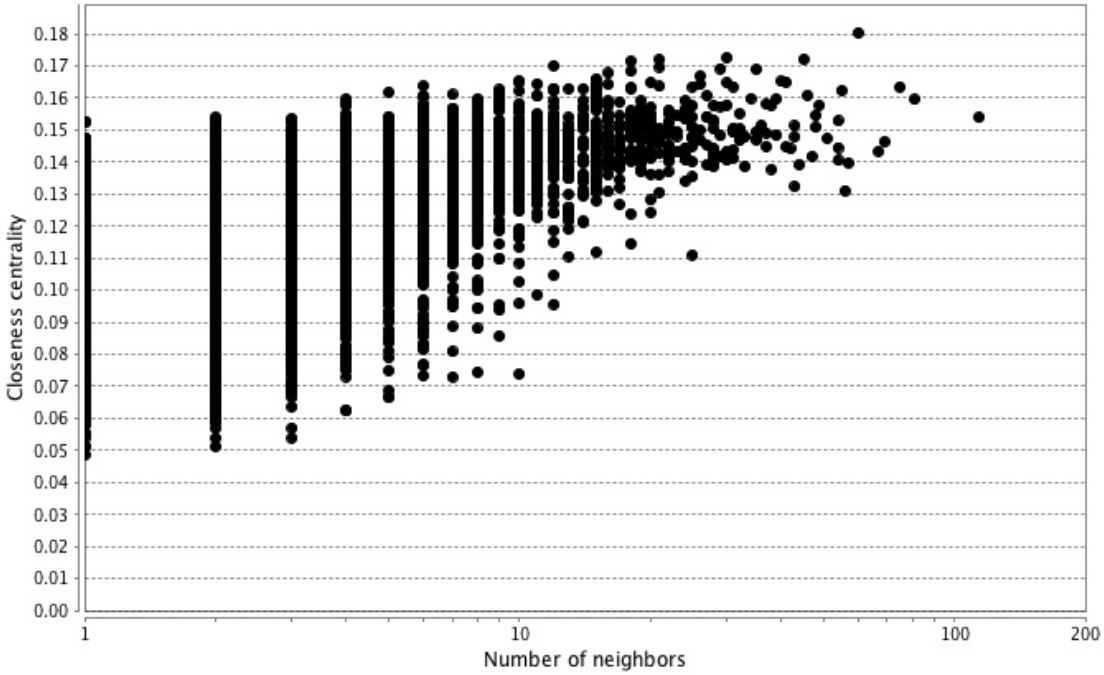


Figure 7. Number of neighbors vs closeness centrality. Closeness centrality is a measure of how close a node is to other nodes or how easy (shortest path) it is to get to another node.

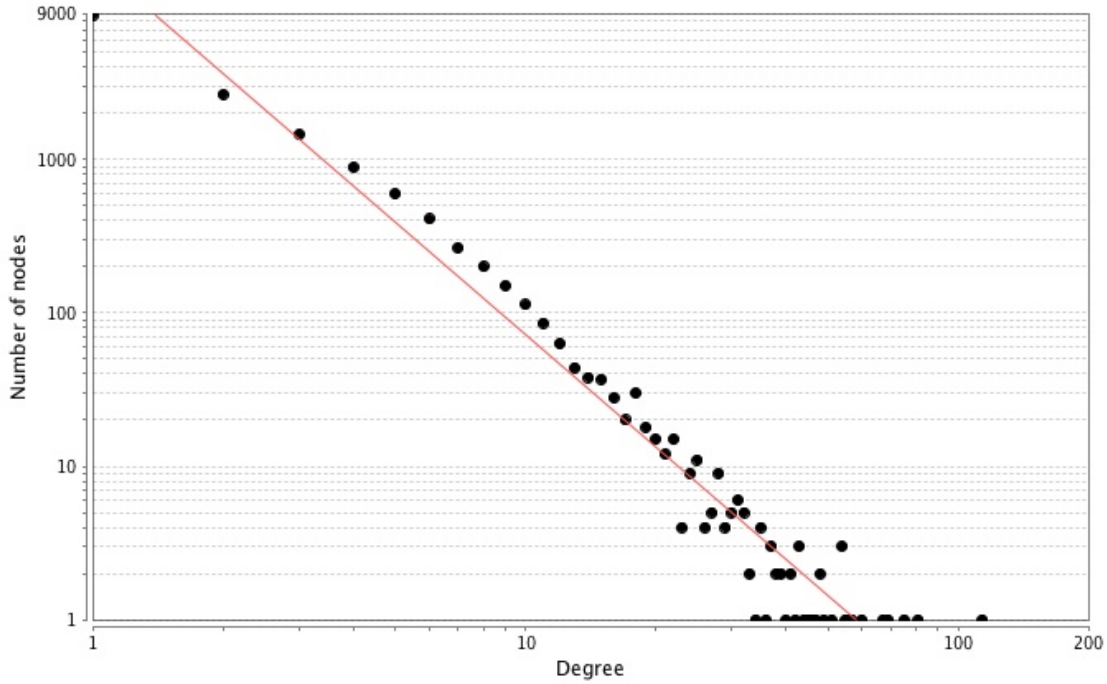


Figure 8. Number of nodes of N-degree. We see that the node degrees follows the power curve extremely well (correlation = 0.987, R-squared = 0.949, $y=19849*x^{-2.434}$) and is very similar to the degree distribution of a random network (see appendix). There are many nodes (~9000) that are only connected to one other node. Also the highest degree node (Rocco Siffredi) has a degree of 114.

Top 10 Nodes by Degree	**Degree**	**Porn or Hollywood Actor/Actress**
Rocco Siffredi	114	Porn actress
Marlene Dietrich	81	Hollywood actress
Lana Turner	75	Hollywood actress
Nikita Denise	69	Porn actress
Nicole Sheridan	67	Porn actress
Warren Beatty	60	Hollywood actor
Randy West	57	Porn actor
Britney Stevens	56	Porn actress
Clark Gable	55	Hollywood actor
Voodoo	54	Porn actor

Table 2. Top nodes by degree and their respective industries.

Top 10 Nodes by 4-Cliques	Number of Cliques	Porn or Hollywood Actor/Actress
Rocco Siffredi	1544	Porn actress
Nicole Sheridan	1520	Porn actress
Nikita Denise	1282	Porn actress
Mark Davis	1054	Porn actor
Lana Turner	1038	Hollywood actress
Marlene Dietrich	986	Hollywood actress
Julian (Julian Rios)	951	Porn actor
Lanny Barbie	807	Porn actress
Taylor Rain	802	Porn actress
Evan Stone	770	Porn actor

Table 3. Top nodes by number of 4-cliques and their respective industries. These would also be known as “partner swaps” where two people share two of the same partners. The total number of 4-cliques for the entire network was 13733.

Top 10 Nodes by 3-Cliques	Number of Cliques	Porn or Hollywood Actor/Actress
Nicole Sheridan	91	Porn actress
Rocco Siffredi	74	Porn actress
Nikita Denise	67	Porn actress
Nina Hartley	44	Porn actress
Voodoo	36	Porn actor
Marlene Dietrich	30	Hollywood actress
Tyler Faith	28	Porn actress
Ginger Lynn Allen	28	Porn actress
Briana Banks	25	Porn actress
Greta Garbo	24	Hollywood actress

Table 4. Top nodes by number of 3-cliques and their respective industries. We notice that all of these people are female and most of them are in porn. This implies that highly sexualized women are likely to sleep with other women. Total number of triangles for the entire network was 542.

Top 10 Nodes by 5-Cliques	Number of Cliques	Porn or Hollywood Actor/Actress
Nicole Sheridan	11434	Porn actress
Rocco Siffredi	10186	Porn actress
Nikita Denise	8861	Porn actress
Mark Davis	5792	Porn actor
Julian	4072	Porn actor
Evan Stone	3961	Porn actor
Voodoo	3280	Porn actor
Marlene Dietrich	2962	Hollywood actress
Jesse Jane	2955	Porn actress
Tyler Faith	2926	Porn actress

Table 5. Top nodes by number of 5-cliques and their respective industries. A 5-cliques (aka pentagon) requires at least one lesbian/gay couple. The porn industry dominates the top pentagons. Total number of pentagons for the entire network was 39875.

Comparison with a Random Network

For proper analysis, we wanted to compare the network with a similar network with randomized links. Each node has the same degree as before but their connections to other nodes were random. To randomize the graph correctly, we found the expected number of links and used a probabilistic mechanism to generate links between nodes. Since our network also has three types of links (relationship, married, encounter), we kept the number of links of a particular type the same as well. Figure 9 shows the distribution of the types of relationships. This random graph was then used for comparison against the actual data. Tables 6-8 show the 4-cliques, 3-cliques, and 5-cliques that came out of the random network. Further analyses of the random network are in the appendix.

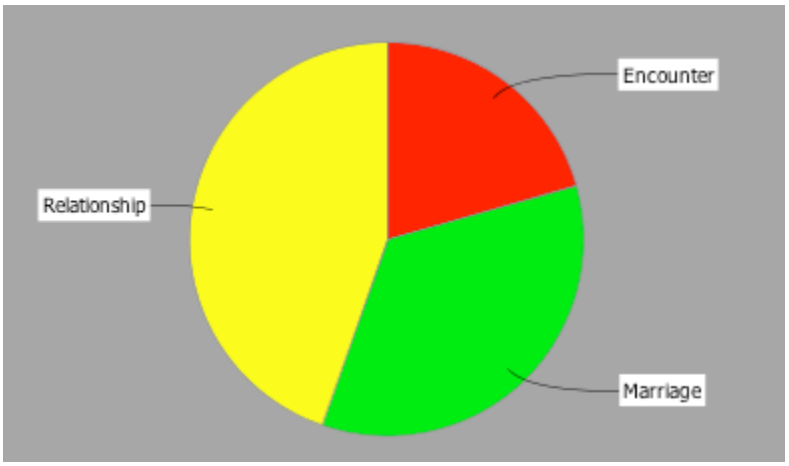


Figure 9. Types of links and their percentage distributions.

Top 10 Nodes by 4-Cliques	Number of Cliques	Porn or Hollywood Actor/Actress
Rocco Siffredi	220	Porn actress
Nikita Denise	146	Porn actress
Britney Stevens	118	Porn actress
Randy West	109	Porn actor
Voodoo	92	Porn actor
Marlene Dietrich	79	Hollywood actress
Lana Turner	62	Hollywood actress
Nicole Sheridan	61	Porn actress
Jesse Jane	59	Porn actress
Hugh M Hefner	51	Porn actor

Table 6. Top nodes by number of 4-cliques and their respective industries. The total number of 4-cliques for the entire network was 1075, much less than 13733 of the actual network. This list from the random network is still dominated by the porn industry.

Top 10 Nodes by 3-Cliques	Number of Cliques	Porn or Hollywood Actor/Actress
Rocco Siffredi	23	Porn actress
Nikita Denise	23	Porn actress
Britney Stevens	13	Porn actress
Jesse Jane	12	Porn actress

Voodoo	9	Porn actor
Nicole Sheridan	9	Porn actress
Randy West	7	Porn actor
Lana Turner	7	Hollywood actress
Daisy Marie	7	Porn actress
Marlene Dietrich	6	Hollywood actress
Karrine Steffans Mccrary	6	Hollywood actress
Frank Sinatra	6	Hollywood actor

Table 7. Top nodes by number of 3-cliques and their respective industries. Total number of triangles for the entire network was 152, compare to 542 from the actual network. This list is no longer completely from the porn industry nor are they all women.

Top Nodes by 5-Cycles	**Number of Cycles**	**Porn or Hollywood Actor/Actress**
Rocco Siffredi	1935	Porn actress
Nikita Denise	1308	Porn actress
Randy West	1140	Porn actor
Britney Stevens	106	Porn actress
Voodoo	760	Porn actor
Marlene Dietrich	703	Hollywood actress
Nicole Sheridan	641	Porn actress
Jesse Jane	600	Porn actress
Hugh M Hefner	492	Porn actor
Lana Turner	482	Hollywood actress

Table 8. Top nodes by number of 5-cliques and their respective industries. Total number of pentagons for the entire network was 8011, much less compared to 39875 5-cliques from the actual network.

Automatic Assignment of Male and Female Nodes

The data we initially collected did not have information about the sex of the node so we decided to predict the sexes using eigenvectors. The probability of the sex of a particular node being the opposite of its neighbors is extremely high (~1) since the majority of the couplings are heterosexual. We denote a male as -1 and a female as 1. For a node N that has n male neighbors and m female neighbors, the sex of N can be determined as -(m-n)/(m+n). This applies a linear transform to the node values that also keeps their values the same. Thus the node values are in the eigenvector associated to the linear transform matrix. Table 6 shows the predicted genders of a few nodes.

Node Name	**Predicted Gender**	**Actual Gender**
Jack Kelly	Male	Male
Jo Anne Worley	Female	Female
Rhonda Worthey	Female	Female
Marie Whitney	Female	Female
D J Guthrie	Male	Male
Charlie Sheen	Male	Male
Mark Dymond	Male	Male

Table 9.

Once we had the predicted genders, we used an online knowledge base called Freebase (http://www.freebase.com/) to gather more information on our nodes, including genders. Just like the table above, our predictions were extremely accurate.

Sexually Transmitted Diseases in Our Network

Since the Hollywood and porn sex network is so highly connected, we thought it would

be interesting to see if we could model the propagation of sexually transmitted diseases. This required finding celebrities with confirmed STDs, dates of when they were confirmed with the disease, and transmission rate of the STDs. Since this information was difficult to find (particularly dates the diseases were confirmed), we narrowed our search down to herpes, which is the most common (at least, in public) STD. Table 10 shows the transmission rates of herpes from person to person. To simplify the calculations, we assumed no condom usage on all links (even though condom usage is 16.1%). If we did not have a node's age, we assume that their age is the average age of the network, which was 47.7 years old.

STD	**Transmission Rate**	**Comments**
Herpes	10%	From male to female, no condom usage
Herpes	4%	From female to male, no condom usage

Table 10.

We found 15 celebrities with confirmed herpes and the dates of their diseases. Starting with these nodes, we calculate the probability of the rest of the network having herpes and when those nodes were expected to be infected. We keep in mind that transmission only affects relationships that occur after the confirmed date of disease. Appendix B lists all 7000 nodes that have a chance of infection, ranked by probability having herpes. Looking at the list we generated, we see that Ashley Olsen (of the Olsen twins) has a 97% chance of having herpes. We expect her to have received the disease in 2005.

Conclusions and Future Work

We notice that the number of triangles (3-cliques) and pentagons (5-cliques) in the Hollywood sex network is much higher than what we would expect from a randomized network with same degree distribution. This makes sense as celebrities (particularly porn stars but also some Hollywood stars like Marlene Dietrich) have sex with a gay/lesbian partner more frequently than in a random network. The number of partner swaps (4-cliques) is also higher for the sex network.

Using Charlie Sheen as our seed proved very fruitful as our network included famous non-actors such as Adolf Hitler and John F. Kennedy. We hypothesize that we reached the majority of the actors in both Hollywood and the porn industry. Next steps would be to find correlations between this sex network and co-actor network from IMDB, and draw conclusions about probabilities of having sex based on whether actors worked together or not. We would also like to separate the Hollywood and porn networks from each other and analyze them individually. This way we could find the key nodes that bridge Hollywood stars to porn stars.

We would like to recalculate the STD propagation, and this time include the rate of condom and other contraceptives usage, the effectiveness rate of the contraceptives in preventing STD transmission, and other STDs besides herpes.

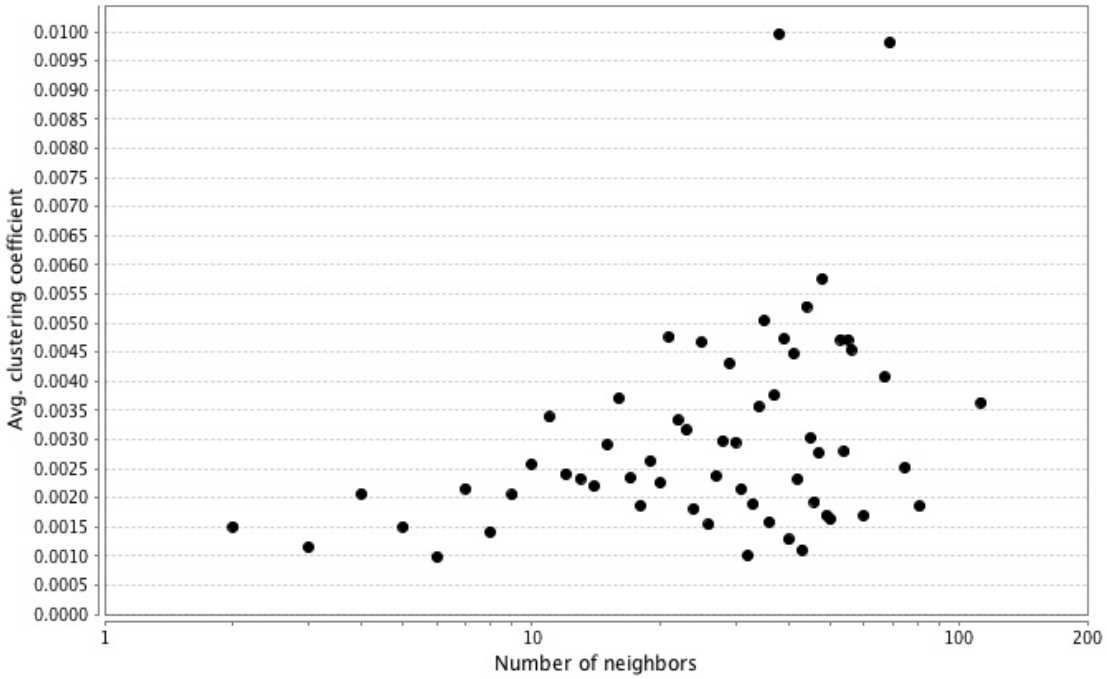
References

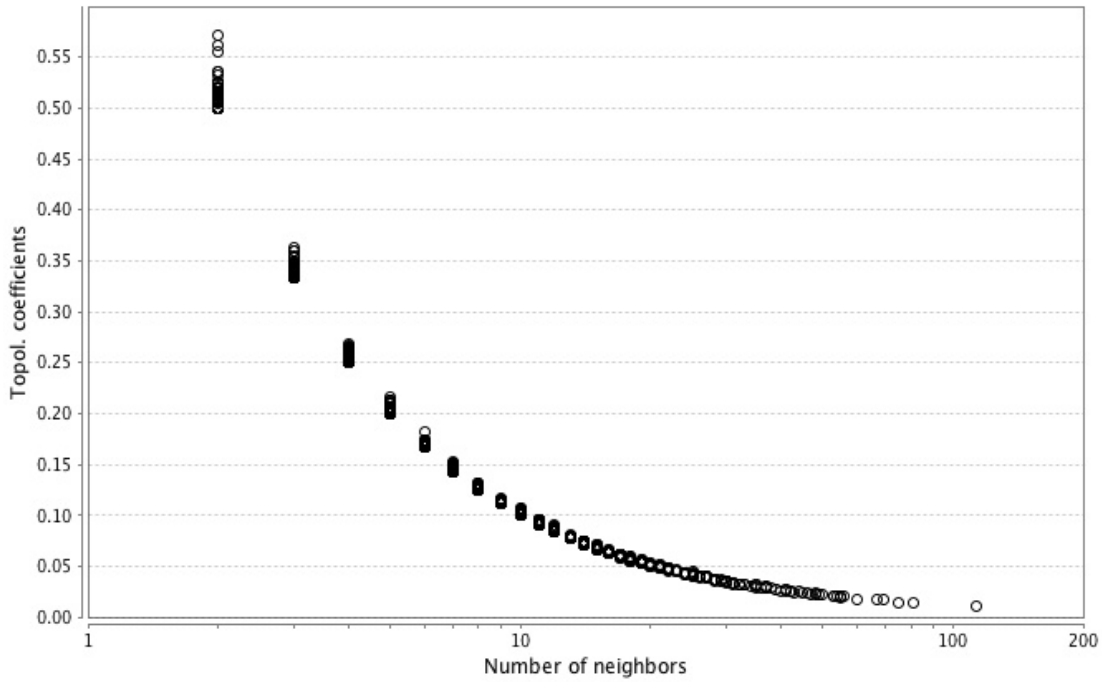
1. http://www.whosdatedwho.com/
2. http://www.freebase.com/
3. http://www.webmd.com/sex-relationships/all-about-herpes

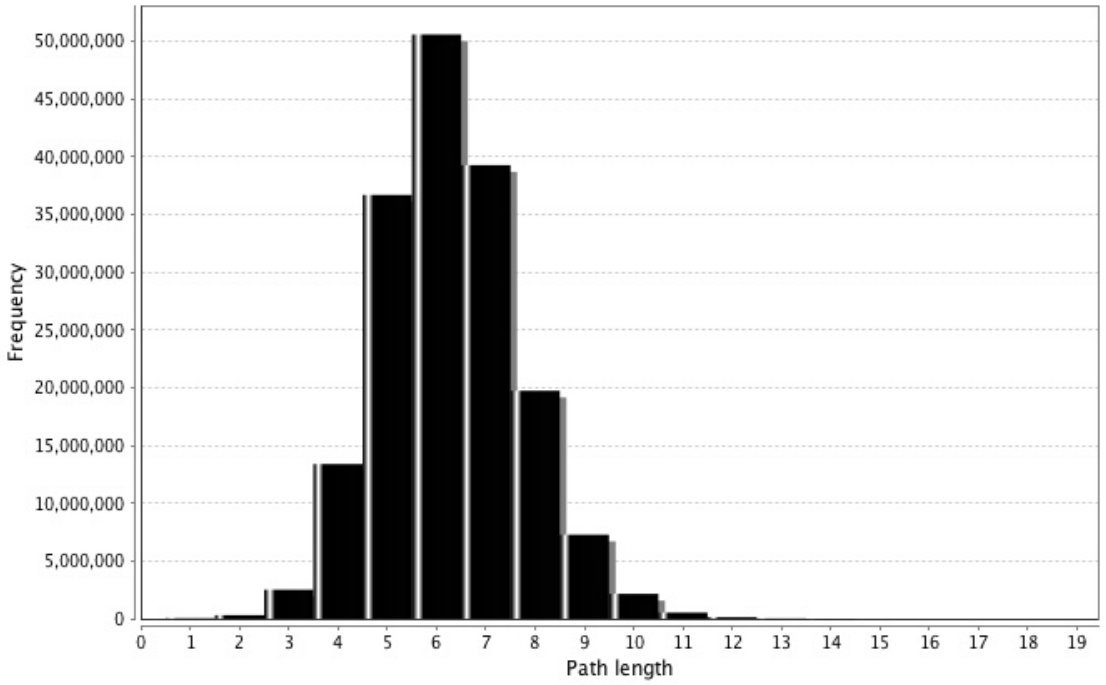
4. http://www.herpesonline.org/std/celebrities-with-herpes/
5. http://www.herpes-coldsores.com/celebrities-with-herpes.html
6. http://www.guttmacher.org/pubs/fb_contr_use.html
7. http://en.wikipedia.org/wiki/Comparison_of_birth_control_methods#Comparison_table
8. http://stdcarriers.com/registry/bio/702-paris-hilton-genitalherpes.aspx
9. http://www.covenanteyes.com/2008/10/28/ex-porn-star-tells-the-truth-about-the-porn-industry
10. http://www.boston.com/ae/celebrity/articles/2007/08/10/jessica_alba_herpes_claim/
11. http://retardzone.com/2008/09/25/top-10-hottest-celebrities-infected-with-stds/
12. http://www.iub.edu/~kinsey/resources/FAQ.html#frequency
13. Sex Degrees of Separation, The Ultimate Guide to Celebrity Relationships. Irad Eyal.

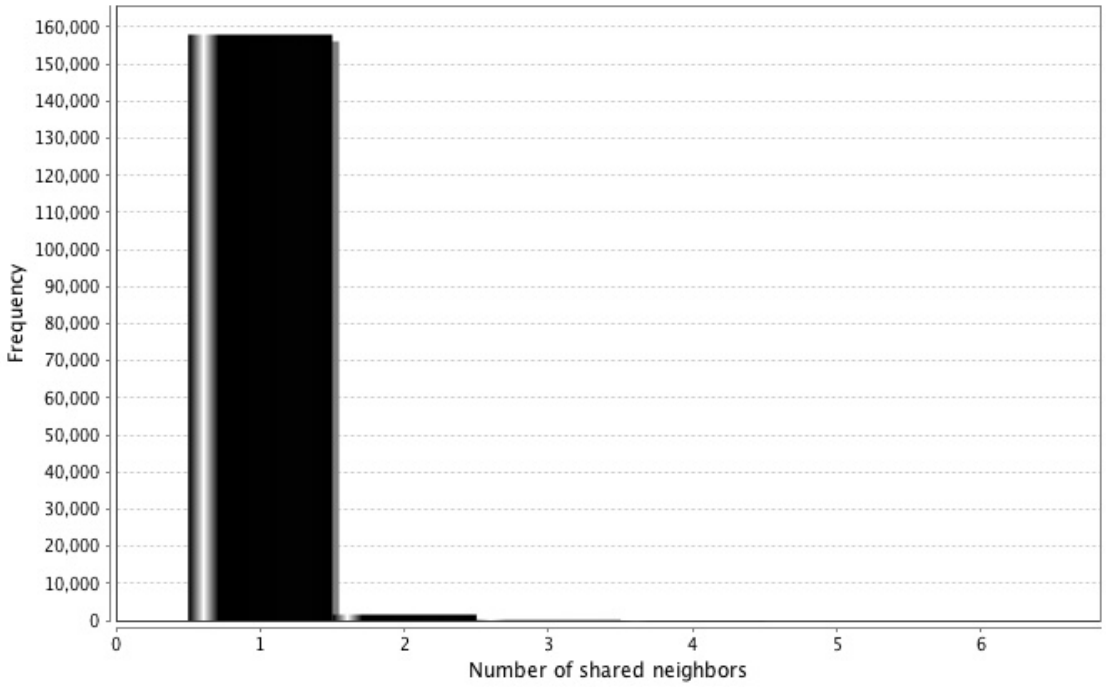
Appendix A

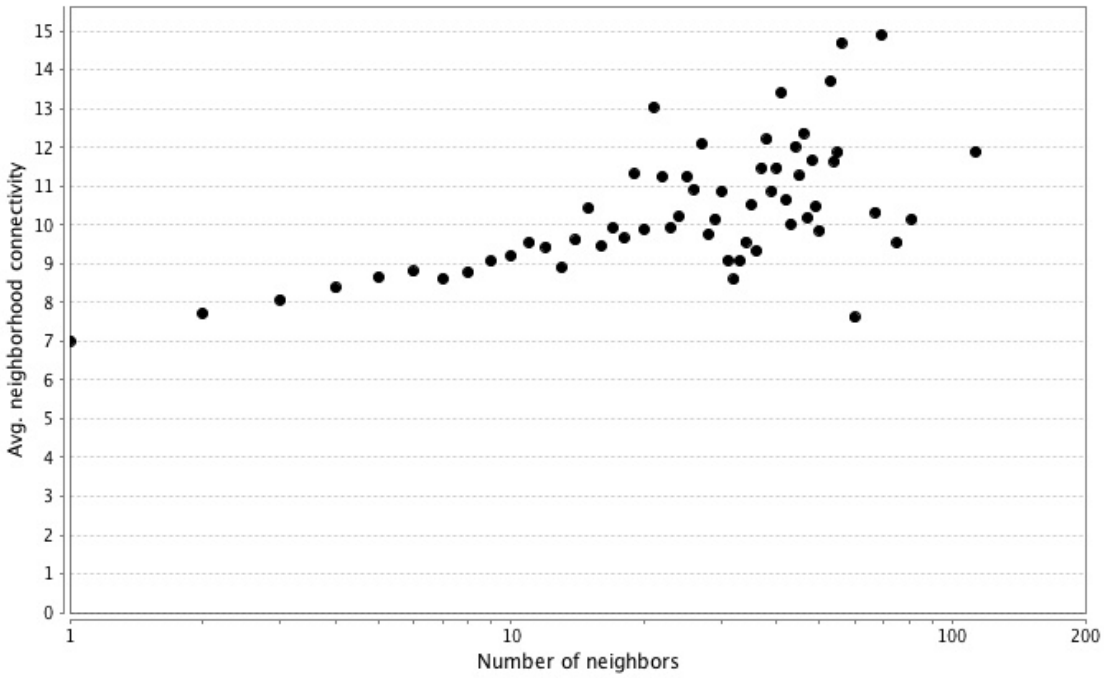
Network information for the randomized network.

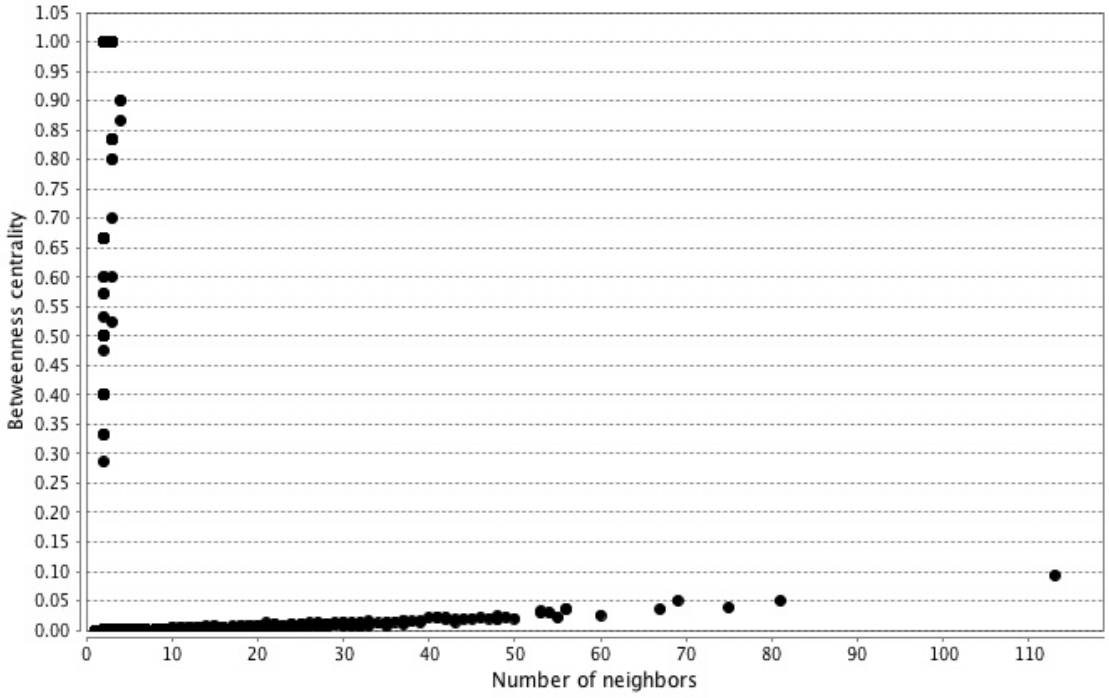


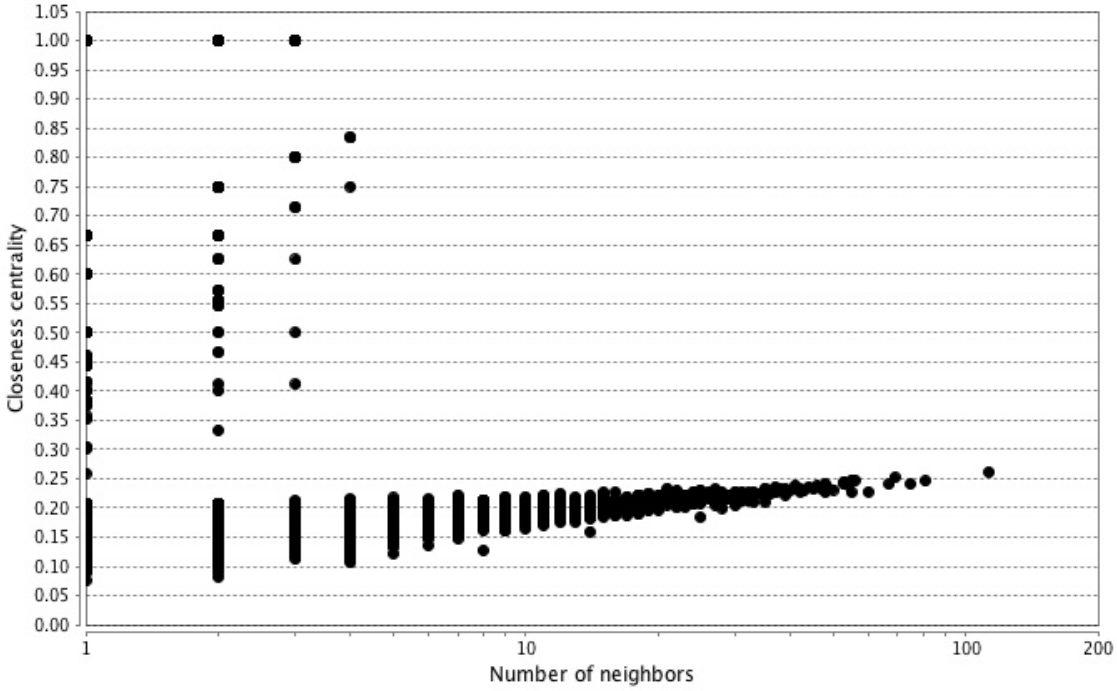


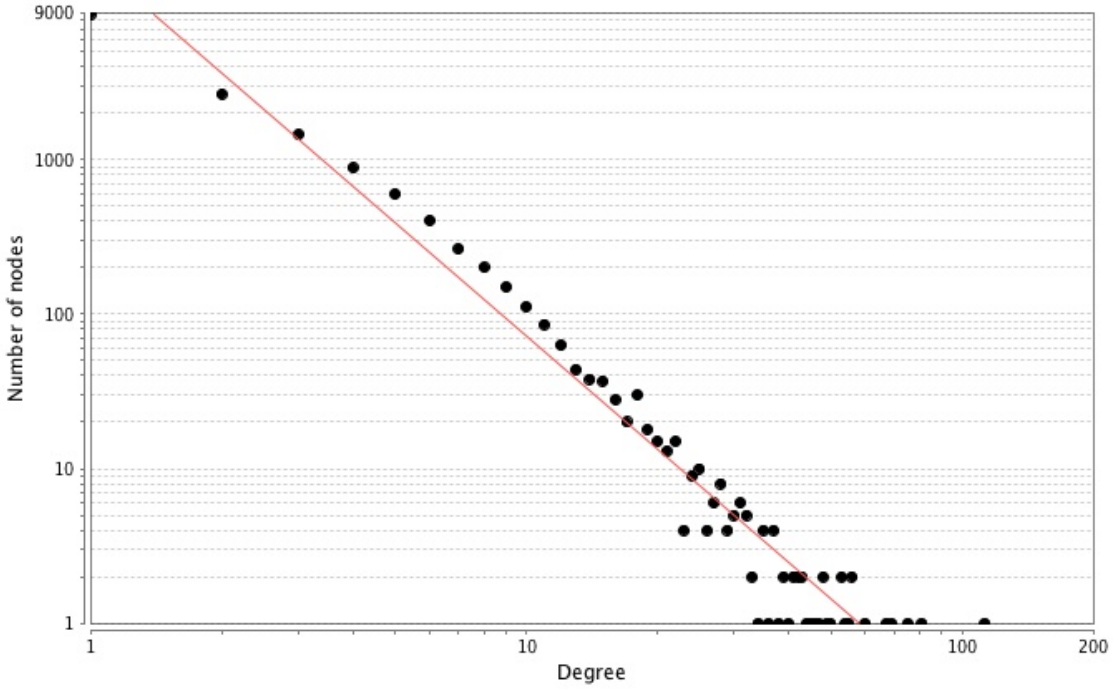

Frequency
50,000,000
45,000,000
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Path length


Frequency
160,000
150,000
140,000
130,000
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
0 1 2 3 4 5 6
Number of shared neighbors


Avg. neighborhood connectivity
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1
10
100
200
Number of neighbors


Betweenness centrality
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
1.05
0
10
20
30
40
50
60
70
80
90
100
110
Number of neighbors





correlation = 0.987, R-squared = 0.949, $y=19861*x^{-2.434}$