CSE 316 Final Project Part 2:

A Study of the Citation Network

Brian Gauch

April 29, 2014

1 Introduction

Social network analysis can be used to study relationships within collections of research literature, and historically this has been an area of much interest. Networks may be built in which authors are nodes and co-authorships are edges. These types of social networks can be used to identify the most influential researchers. Other networks can be constructed in which papers are the nodes and citations are the edges. These networks can be explored to identify key papers within a discipline. Citations have been used to explore document collections long before the more current research on text-based social networks to support communities of scholars.

This project explores three key ideas from social network analysis applied to a network of documents linked via their citations. We used data about High-energy physics citation network, available from Stanford's Large Network Dataset Collection. The dataset contained roughly 35,000 papers (nodes) and 420,000 citations (edges). Each paper had its date of publication, but we were not provided the paper titles, authors, contents, or other data.

We tested to see if the citation network is truly scale-free, as had been previously suggested [5]. Second, we explored whether or not older papers are less likely to get cited today. Finally, we studied the overall structure of the citation network, in particular looking for evidence of homophily in fields and sub-fields.

2 Key Ideas

2.1 Scale Free – or is it?

If we think of citations of a paper as indicating that paper's popularity and/or intrinsic value, then we expect new papers to take this into account, and be more likely to cite previously-cited papers. If it is based on popularity, this is called "preferential attachment", and this rich-get-richer behavior results in a "Power Law" distribution. A Power Law distribution has a number of nice properties, foremost of which is that it suggests that we can model it as a scale-free network. A scale-free network is so called because it looks the same regardless of its size, if you zoom appropriately of course.

Experiment 1.1: Plotting the in-degree of the true network, Figure 1.1

Experiment 1.2: Plotting the out-degree of the true network, Figure 1.2

To examine whether or not our graph followed the Power Law, we calculated the in-degree (Expt. 1.1) and the out-degree (Expt 1.2) of the citation network.

Experiment 1.3: plotting generated Barabási–Albert network with m=12 (degree) , Figure 1.1

We used built-in NetworkX graph generators to create networks with in-degrees that followed the Barabási–Albert model, a model that exhibits the Power Law.

Experiment 1.4: plotting generated Erdos Renyi network with p such that m~12 (degree) , Figure 1.1

We used built-in NetworkX graph generators to create networks using the Erdos Renyi model, a model that exhibits random (roughly normal) in-degrees.

Figures 1.0 compares the in-degree distribution of the citation network to the Barabási–Albert and a random distribution. From this, we can see that the in-degree seems to closely follow the Power Law. Figure 1.1 also shows the out-degree compared to the Power Law and it seems to follow the Power Law somewhat less closely.

Experiment 1.5: plotting generated gn_semi_preferential networks (in degree) , Figure 1.2

In the final experiment, we generalized the Barabási–Albert model of preferential attachment to create a directed network. The links are selected with probability proportional to that nodes popularity (number of preexisting links) plus some base uniform probability. In the classic Barabási–Albert model, which is implemented in NetworkX, the number of undirected edges a node has is used as its weight, which serves to give each node a base weight of m (here by weight we mean unnormalized selection probability). However, we return a undirected graph as a result of using this trick, and since we are primarily concerned with the number of citations a paper gets, not the number of times it cites, this is not ideal.

We have reimplemented the Barabási–Albert model more generally in gn_semi_preferential, which uses NetworkX's gn_graph to do most of the heavy lifting. gn_semi_preferential with parameter pref = 1.0 generates a directed Barabási–Albert graph. gn_semi_preferential with pref = 0.0 generates a graph by adding nodes sequentially and linking each node uniformly at random, so old nodes will still have more connections. We got the same distribution of node degrees as the NetworkX implementation when we treated our edges as undirected. However, because of the other NetworkX function we are building on top of, gn_graph, our generalization currently only works with m = 1.

Figure 1.1 compares our directed network that better models the way citations work to the undirected network. If the citation network is truly scale-free, as was long ago suggested[5], and follows a Power Law distribution, the log-log plot of node degrees should be linear. However, we see that the distribution is in fact slightly concave on a log-log plot. We were not able to generate networks that fully mimicked this behavior, but we did get some concavity just by looking at only in-degree.

2.2 Studying citation distribution

We can study the citation network to find out more about citation patterns and also study citation trends over time. Our first experiments gathered data to better understand the overall citation statistics. They are summarized below:

Experiment 2.1: Average citations

The mean number of citations is merely the mean in-degree (or the mean out-degree), which we found to be 12.20. The median in-degree was 4 and the median out-degree was 8. This indicates that there is a small subset of highly-cited papers, which is to be expected. (Figure 2.1)

Experiment 2.2: Number of uncited papers

There were 6,316 papers with an in-degree of 0, indicating that there was a very large number of papers that were never cited at all. This represents 18.3% of all papers in the collection. (Figure 2.1)

Experiment 2.3: Number of self-citations

Surprisingly, there are 44 self-loops – cases where a paper cited itself! This must be due to some cheeky authors. (Figure 2.1)

Experiment 2.4: Number of citation loops

There are also, surprisingly, 657 reciprocated edges, which should not be possible if one can only cite other published papers. This could be due to authors citing some of their unfinished work, or a friend's unfinished work, which was later published or was published simultaneously. These are fairly uncommon happenings considering that there are 34,546 nodes in our data. (Figure 2.1)

Experiment 2.5: Paper citations over time

We expected that the number of times a given paper is cited would decrease over time. We had to be careful when searching for this effect; we needed to take into account the increasing total number of citations per year, and the increasing size of the pool of citable papers. Thus, we compared the citations over time in our citation graph G to a random graph, T, which is a copy of G in which the publication dates were scrambled. Thus, T has the same size as G, and the same distributions of degree and date – just no correlation between dates and edges. For each graph, for each edge, we extracted the time gap between the citing node and the cited node to calculate the elapsed time between a paper's publication date and its citation date(s).

Figure 2.2 shows the distribution of edge date differentials for T and G. In our true graph, G, we found that the popularity of a paper generally increases for the first 100 days or so, then decays exponentially. The random graph T had only linear decay, and no peak in popularity after initial publication. We get this distribution for T because we are essentially selecting points at random from the interval [date of first publication, date of last publication].

2.3 Fields and subfields

This citation data is over a long time period and from different fields, such as computer science, bioinformatics, math and physics. We expect to see heavy homophily based on a paper's research area, so the network should be broken up into communities. Unfortunately, communities are difficult to detect, and the only community-related method that NetworkX provides detects k-clique communities, which is slow and does not give us a full picture.

We attempted to glean more information about the network structure by looking at betweenness (sometimes used to break networks up into communities), clustering, and connected components. For comparison, we generated a Barabasi-Albert graph with roughly the same number of nodes and edges as the citation network. We shall refer to the citation network as G, and the generated test network as T.

Experiment 3.1: finding mean clustering in G and T, Figure 3.1

Because the collection of papers were drawn from math and science, there was effectively one connected component, with few tiny separate components. In particular, 34,401 of the 34546 nodes were in the main component, with other components ranging in size from 2 to 6. In contrast, T contained only one component, as an artifact of how it was created.

Experiment 3.2: finding connected components in G and T, Figure 3.1

G had clustering coefficient 0.2848 whereas T's clustering coefficient was a mere 0.0048, i.e., G had ~59 times as many closed triangles. Thus, the citation network was far more clustered than the T graph, which is a good indicator of homophily on the small scale (related papers, not whole related fields).

Experiment 3.3: finding k-clique communities in G and T, Figure 3.1

There were nineteen 15-clique communities in G but none in T. T's largest k-clique communities generally had k of 5. The k-clique community data hinted at larger scale homophily, because much larger cliques and clique communities formed in G.

Experiment 3.4: finding mean betweenness in G and T, Figure 3.1

G had a higher mean betweenness (1.4x10-4) compared to T (6.42x10-5). Higher mean betweenness for G implies longer shortest-path distances between two nodes. This makes sense, assuming there is more homophily in G, since in our randomly generated graph, with no homophily, we would expect the small-world phenomenon (short distances).

Experiment 3.5: finding median betweenness in G and T, Figure 3.1

G had a lower median betweenness (1.9x10-8) compared to T (8.3x10-6).The difference in mean and median betweenness suggests that there are some nodes with high centrality, some major papers that many researchers cite.

Experiment 3.6: plotting distribution of betweenness in G and T, Figure 3.2

The G network also showed significantly more variance in its distribution of betweenness, indicating a somewhat hierarchical structure.

3 Conclusions

We found that the citation network was roughly scale-free, particularly the distribution of degree for popular nodes. The in-degree distribution seemed to follow the Power Law more closely than the out-degree. The shape of the distribution for unpopular nodes proved difficult to account for, but could be due to lessening popularity effects. We implemented a directed Barabási–Albert graph generator in order to better model citations and better understand the deviations from the Power Law.

We found that the mean number of citations was 12.2, however the median number of citations 8 (mean out-degree) and the mean number of citations was 4 (mean in-degree). Thus, paper citations have a very non-uniform distribution. We found that papers gain popularity for the first 100 days after their publication, and then their popularity decays exponentially. It is important to keep in mind that while this could be due to popularity effects, but it could also be due to any number of other things such as decreasing relevance.

We were unable to find any hard evidence for the existence of fields and subfields, and homophily therein. There was one large connected component that contained essentially all of the nodes in the network. In addition, the graph had a much higher clustering coefficient than a random network, indicating homophily. Although we did not find disjoint subfields, large scale homophily was indicated by the existence of k-clique communities with k=15. Since the variation of betweenness was higher in our citation network G than in a comparison random network, it is likely that G has a more hierarchical structure than a random Barabási–Albert graph, providing evidence of a somewhat hierarchical structure.

4 Appendix

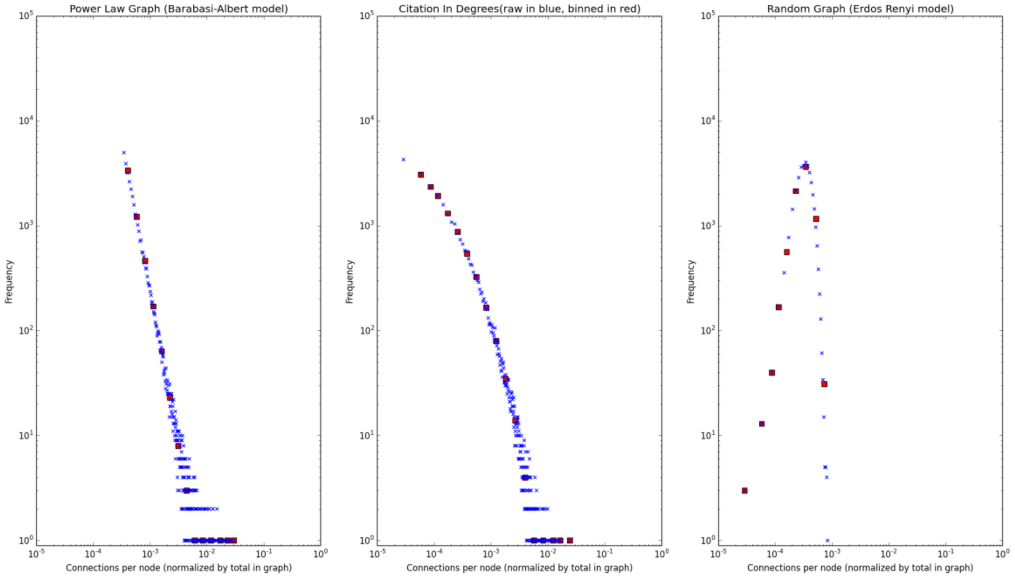


Figure 1.1: In Degree and Power Law

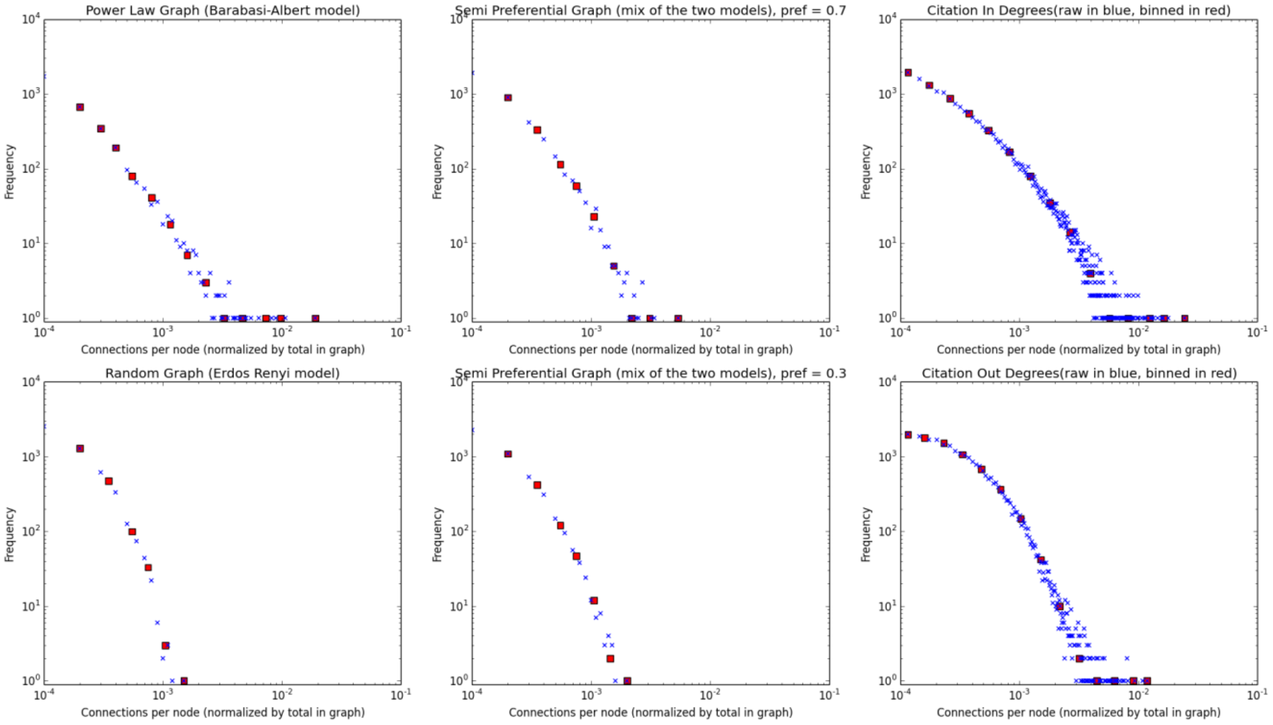
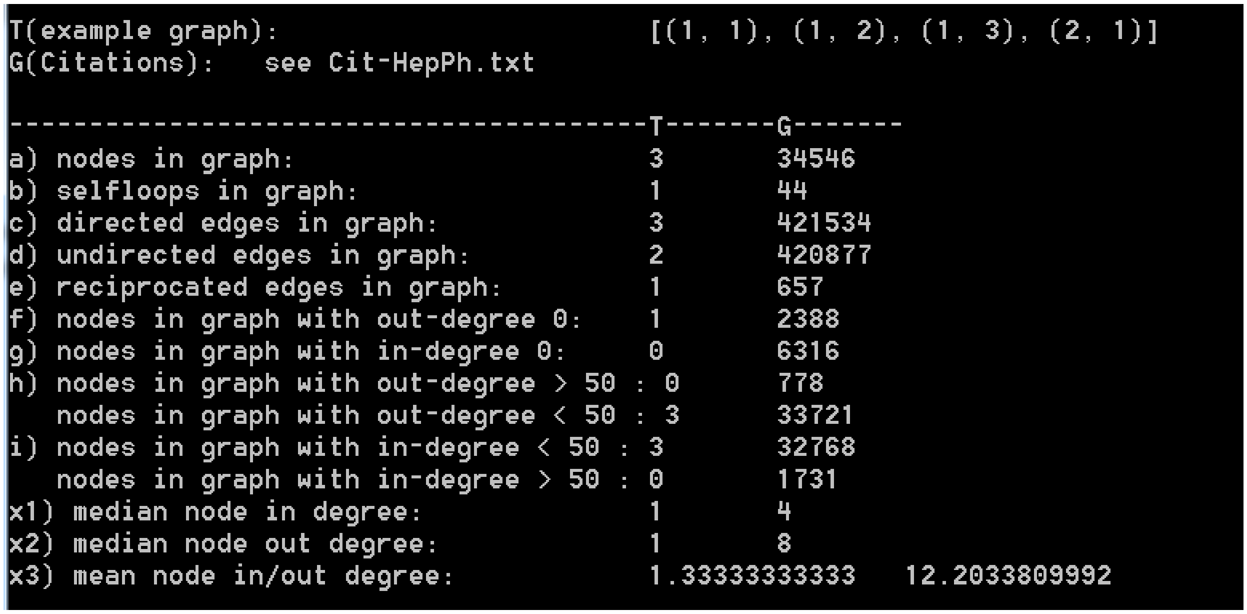


Figure 1.2: Barabasi-Albert Generalization

Unfortunately, out degree=m=1 for all of these generated networks, but we can see the slight curvature of the “Semi Preferential” generalized Barabasi-Albert**,** closer to the shape of the citation in-degree distribution.



```
T(example graph):                       [(1, 1), (1, 2), (1, 3), (2, 1)]
G(Citations):   see Cit-HepPh.txt

----------------------------------------T-------G-------
a) nodes in graph:                      3       34546
b) selfloops in graph:                  1       44
c) directed edges in graph:             3       421534
d) undirected edges in graph:           2       420877
e) reciprocated edges in graph:         1       657
f) nodes in graph with out-degree 0:    1       2388
g) nodes in graph with in-degree 0:     0       6316
h) nodes in graph with out-degree > 50 : 0      778
   nodes in graph with out-degree < 50 : 3      33721
i) nodes in graph with in-degree < 50 : 3       32768
   nodes in graph with in-degree > 50 : 0       1731
x1) median node in degree:              1       4
x2) median node out degree:             1       8
x3) mean node in/out degree:            1.33333333333   12.2033809992
```

Figure 2.1: Screenshot of basic network data

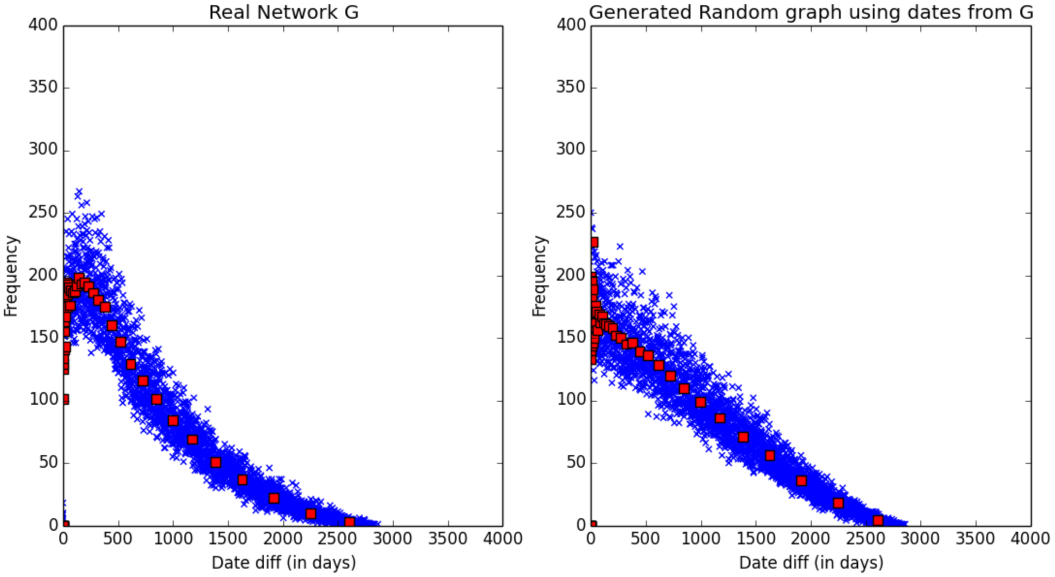


Figure 2.2: Popularity Deterioration

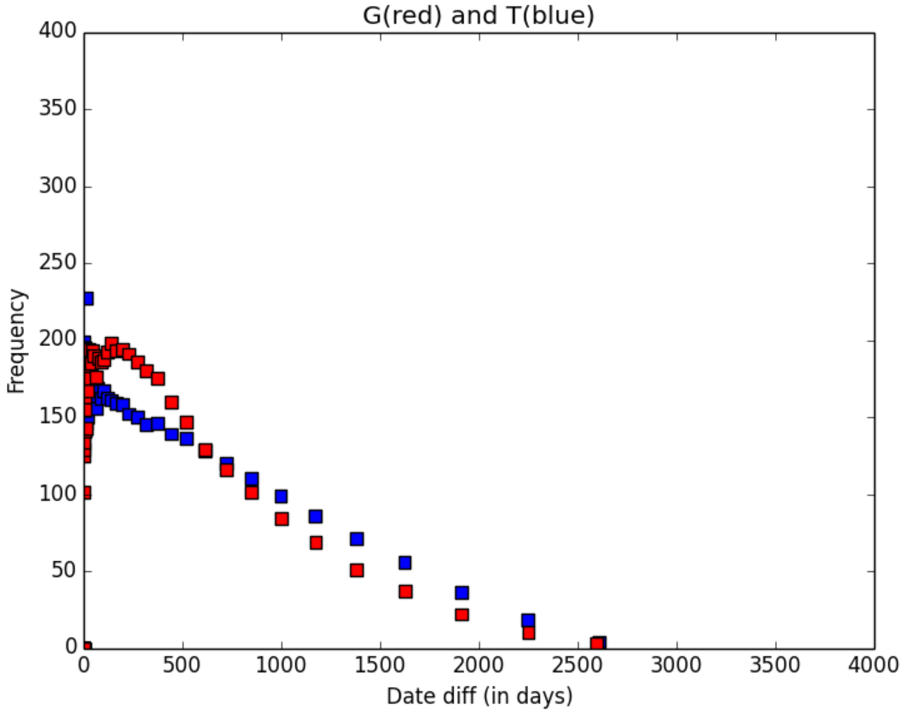
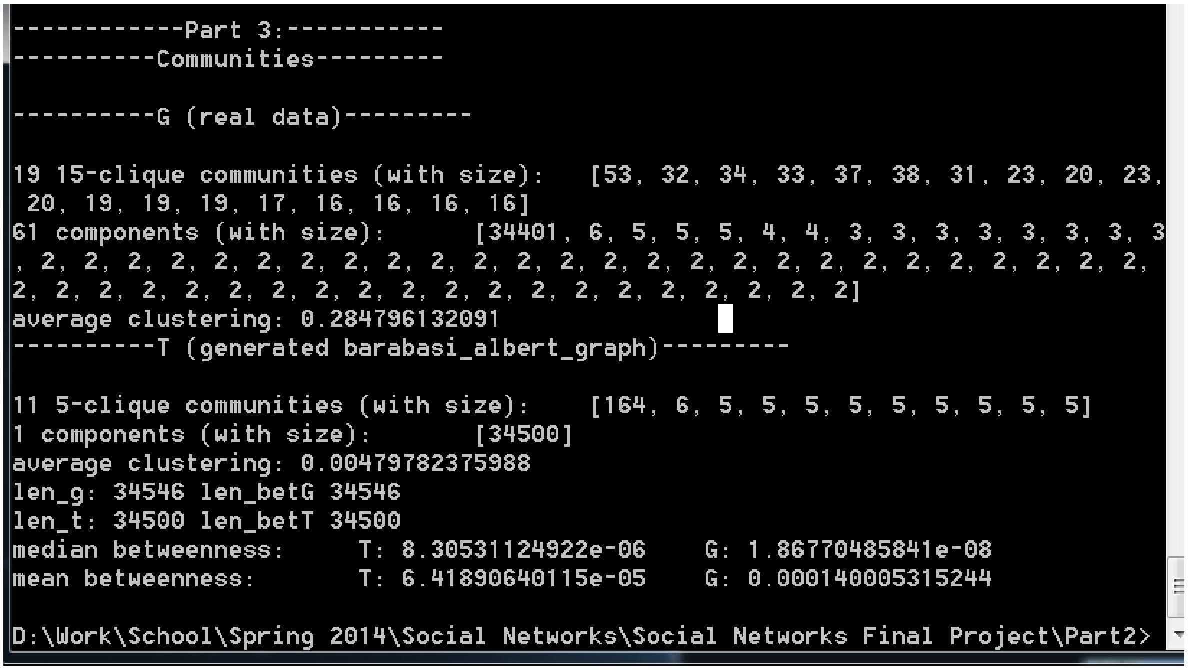


Figure 2.3: Overlay of actual date diff (G) and random date diff (T)



```
------------Part 3:-----------
----------Communities---------

----------G (real data)---------

19 15-clique communities (with size):   [53, 32, 34, 33, 37, 38, 31, 23, 20, 23,
 20, 19, 19, 19, 17, 16, 16, 16, 16]
61 components (with size):      [34401, 6, 5, 5, 5, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3
, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]
average clustering: 0.284796132091
----------T (generated barabasi_albert_graph)---------

11 5-clique communities (with size):    [164, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5]
1 components (with size):       [34500]
average clustering: 0.00479782375988
len_g: 34546 len_betG 34546
len_t: 34500 len_betT 34500
median betweenness:     T: 8.30531124922e-06    G: 1.86770485841e-08
mean betweenness:       T: 6.41890640115e-05    G: 0.000140005315244

D:\Work\School\Spring 2014\Social Networks\Social Networks Final Project\Part2>
```

Figure 3.1: Screenshot of text output

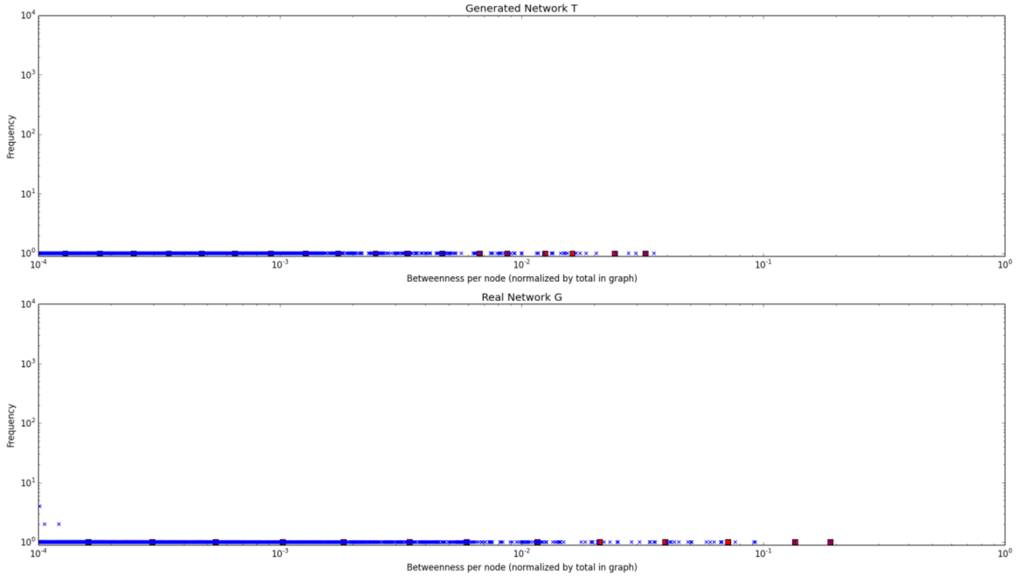


Figure 3.2: Betweenness, approximated with k=100

5 References

[1] Wikipedia contributors, "Scale-free network," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Scale-free_network&oldid=605348610 (accessed April 28, 2014).

[2] Wikipedia contributors, "Power law," *Wikipedia, The Free Encyclopedia,*http://en.wikipedia.org/w/index.php?title=Power_law&oldid=604867862 (accessed April 28, 2014).

[3] Wikipedia contributors, "Barabási–Albert model," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Barab%C3%A1si%E2%80%93Albert_model&oldid=598500765 (accessed April 28, 2014).

[4] Wikipedia contributors, "Preferential attachment," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Preferential_attachment&oldid=588290933 (accessed April 28, 2014).

[5] De Solla Price, D. J. (1965). "Networks of Scientific Papers". Science 149 (3683): 510–515. doi:10.1126/science.149.3683.510. PMID 14325149 (actually found at http://www.garfield.library.upenn.edu/papers/pricenetworks1965.pdf)

[6] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

6 Notes

Thanks to Stanford for the data at http://snap.stanford.edu/data/cit-HepPh.html, the primary data source for this paper.

All coding done in python.

Plots generated with matplotlib.