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. This is somewhat reminiscent of a fractal, although because we are only talking about the distribution,

In one generalization of the Barabási–Albert model of preferential attachment, each new node links to m old nodes. 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.

The other motivation for our implementation of gn\_semi\_preferential, aside from getting a directed version, was to generate networks that more closely approximated the citation 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 Deterioration of relevance/popularity over time**

We expect the number of times a paper is cited to decrease over time. We have to be careful when searching for this effect; we should take into account the increasing total number of citations per year, and the increasing size of the pool of citable papers.

With this in mind, we generated a random graph, T, using the nodes and edges from the true data, G, but with dates selected from G at random. This way, T has the same size as G, and the distributions of degree and date – just no correlation between dates and edges. We then plotted the distribution of edge date differentials for T and 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 method NetworkX provides detects k-clique communities, which takes some time 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. 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.

Unsurprisingly there was effectively one connected component, with some tiny ones. 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). The k-clique community data hinted at larger scale homophily, because much larger cliques and clique communities formed in G. There were several 15-clique communities in G but none in T, only ~5-clique communities.

The G network also showed more significantly more variance in its distribution of betweenness, indicating a somewhat hierarchical structure. Higher mean betweenness for G also implies longer shortest-path distances between two nodes. This makes sense since in our randomly generated graph, with no homophily, we would expect the small-world phenomenon (short distances).

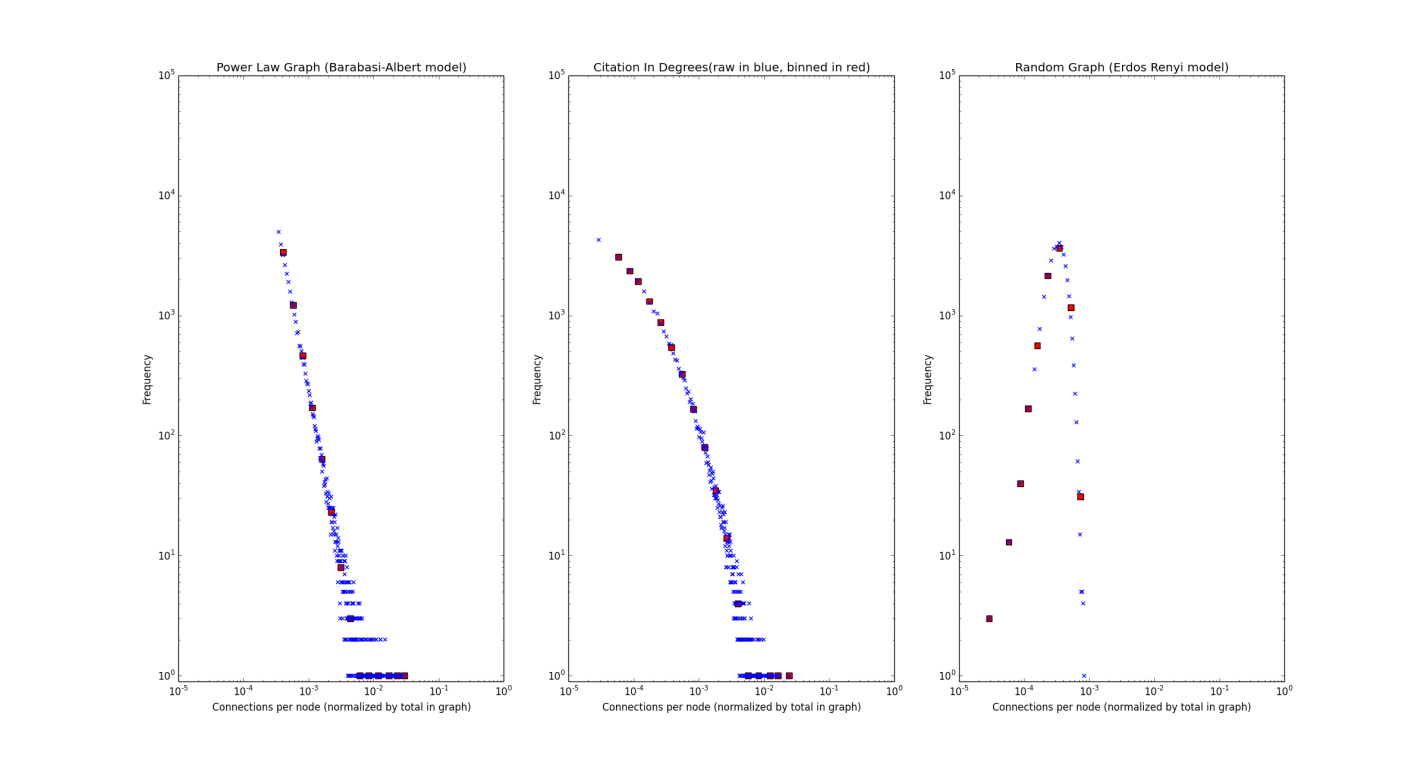
**2.x Other minor findings**

Surprisingly, there are 44 self-loops – cases where a paper cited itself! This must be due to some cheeky authors. 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. The median in degree was 4. The median out degree was 8. The mean degree was 12.20.

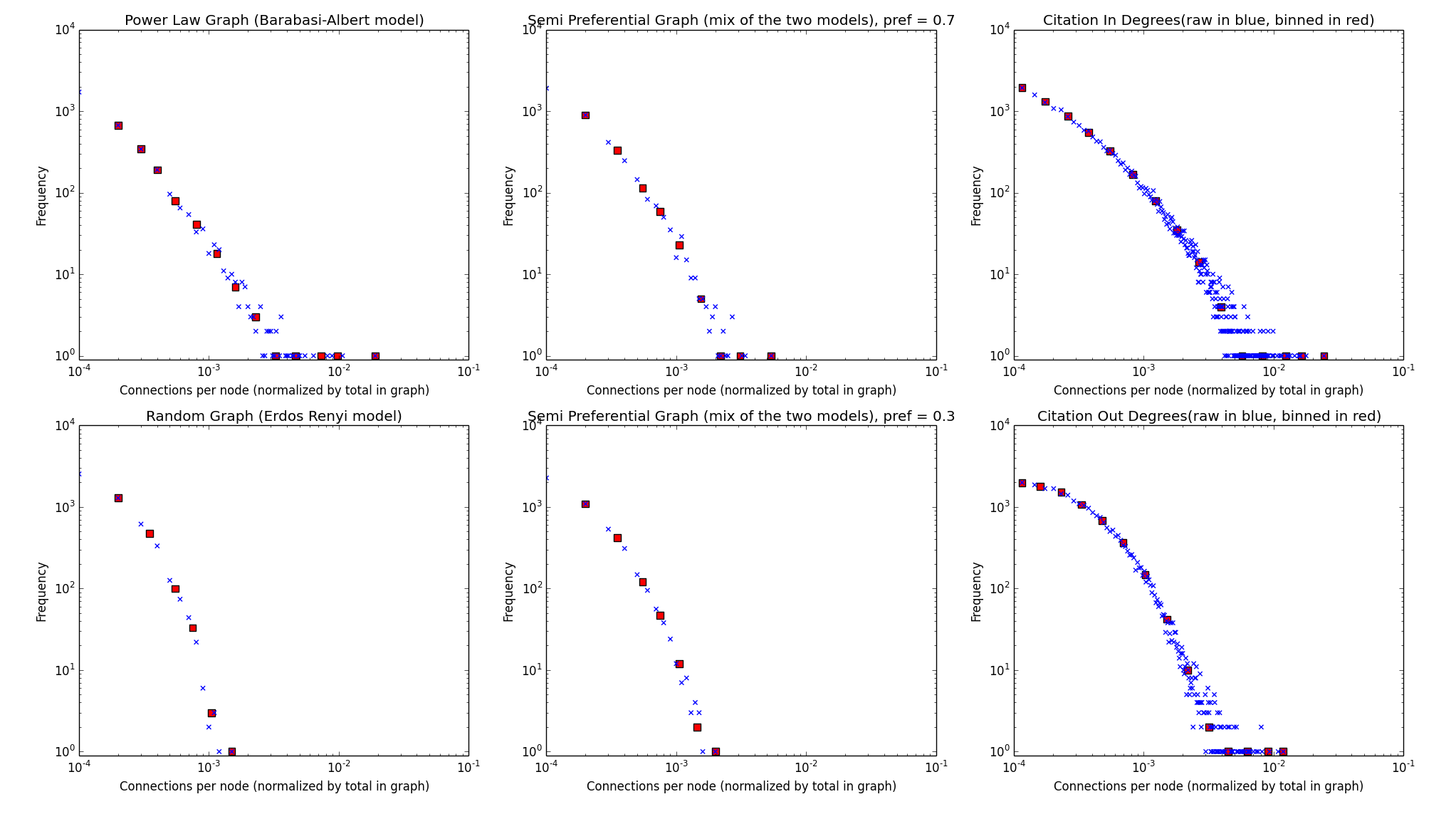
**3 Conclusions**

We found that the citation network was roughly scale-free, particularly the distribution of degree for popular nodes. The shape of the distribution for unpopular nodes proved difficult to account for, but could be due to lessening popularity effects. We found that papers gain popularity for the first 100 days after their publication, and then their popularity decays exponentially. We were unable to find any hard evidence for the existence of fields and subfields, and homophily therein, but we were able to show evidence of local homophily and a somewhat hierarchical structure.

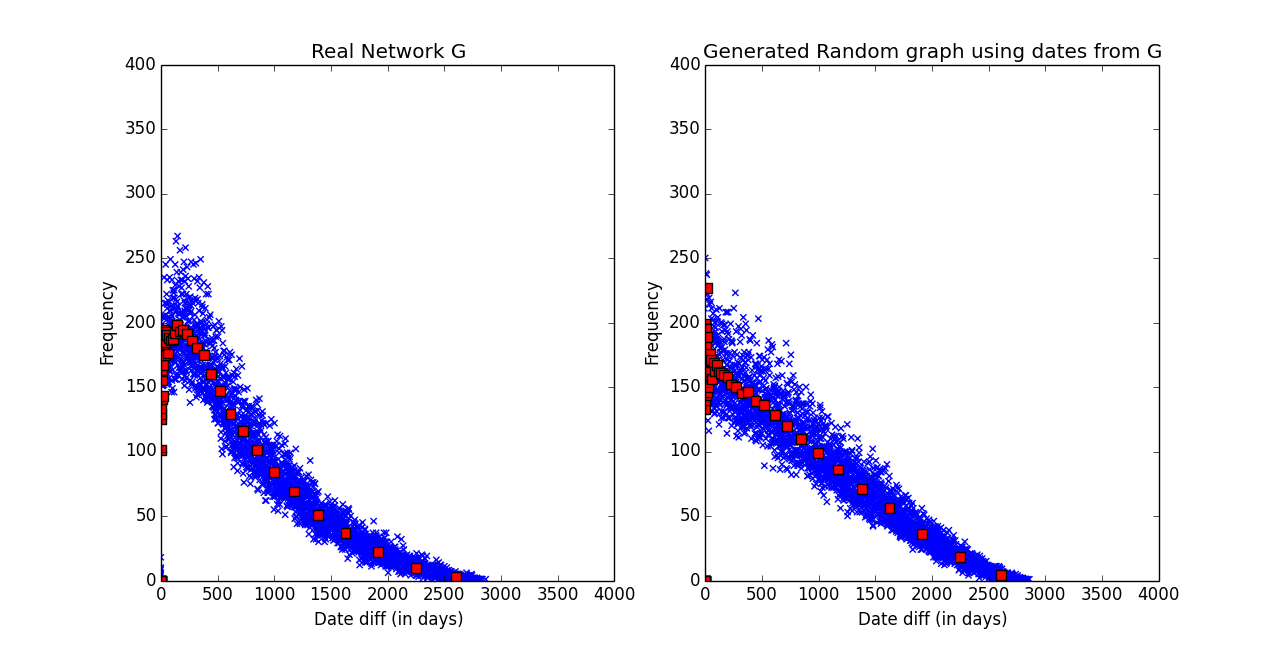
**4 Appendix**

****

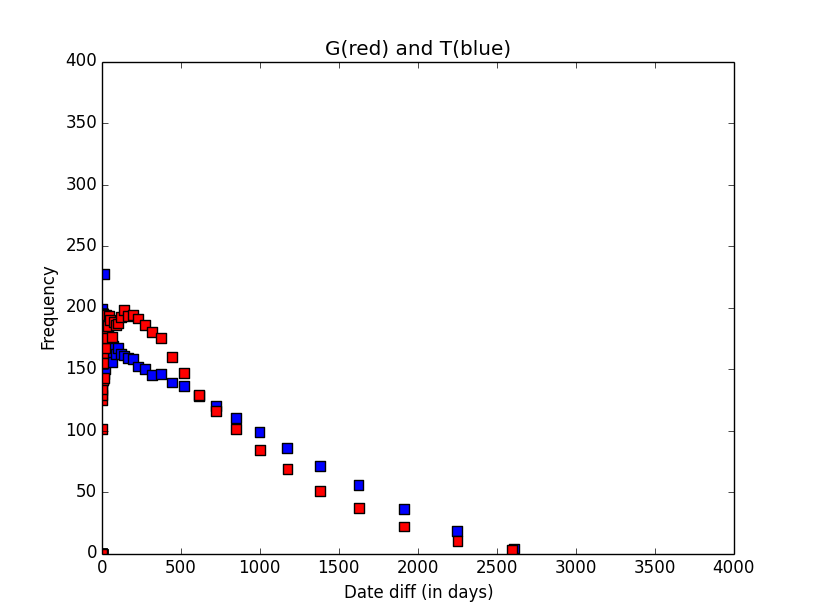
**Figure 1.0: In Degree and Power Law**

**Figure 1.1: Barabasi-Albert Generalization**

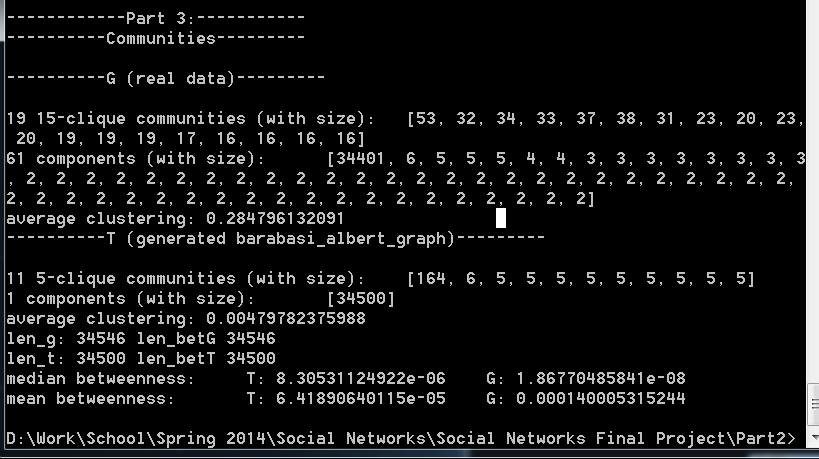
Unfortunately, out degree=m=1 for all of these generated networks, but we can see the slight curvature of the “Semi Preferential” generalized Barabasi-Albert**,** closer to the shape of the citation in-degree distribution.

****

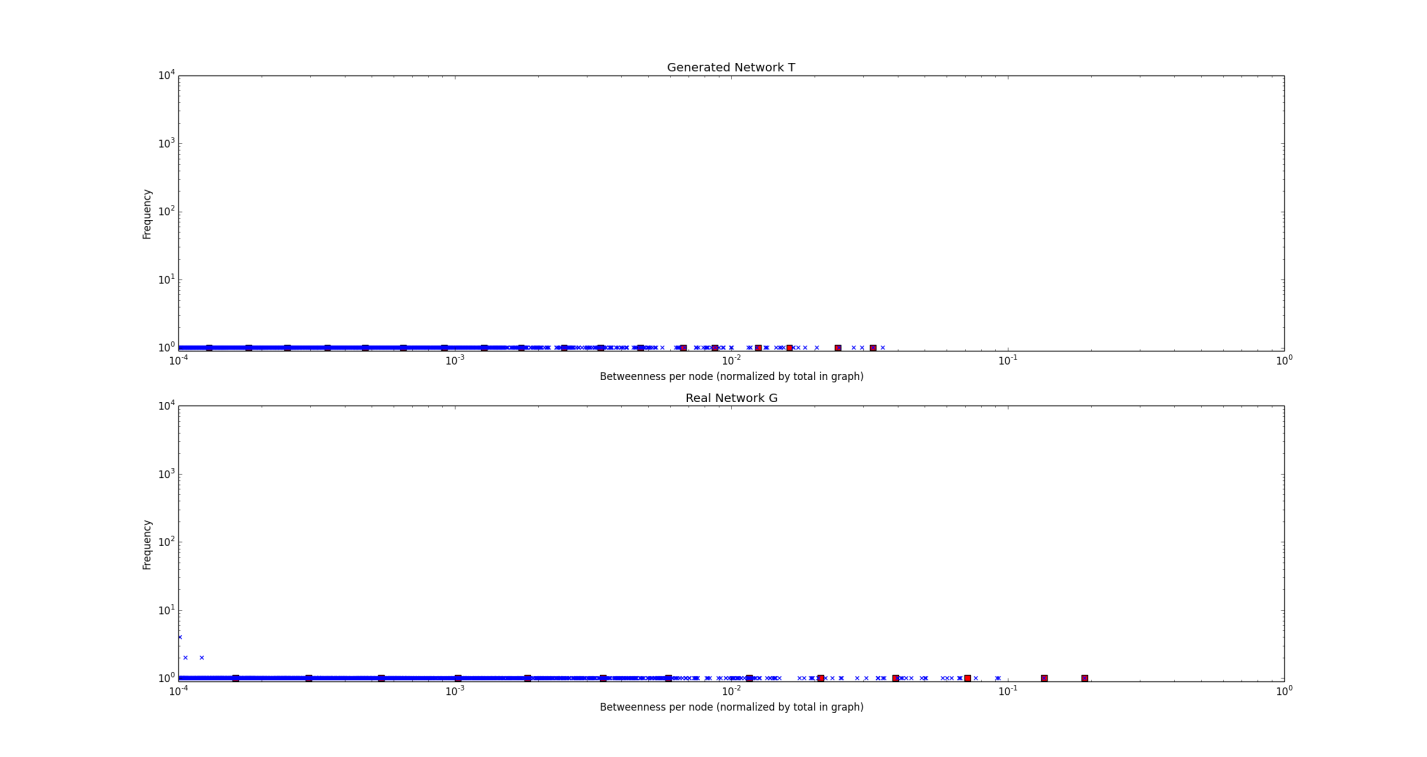
**Figure 2.0: Popularity Deterioration**

****

**Figure 2.1: Overlay of actual date diff (G) and random date diff (T)**

****

**Figure 3.0: Screenshot of text output**



**Figure 3.1: 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](https://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.