Excursion in Social Networks

Jeremy Avnet*

26th January 2003

1 Statistical Measurements

The ability to distinguish things from one another is one of the most basic abilities of any living form. However, the shear number of non-identical things means the ability to consider two things identical is insufficient for an organism needing to simply model its world. By classifying similar characteristics into more general qualities, an organism can make coarse grain comparisons.

Given two undirected graphs, how can we say whether these graphs are the same, not in a sense of them being exactly indentical, but in sense that they express similar qualities? Whenever we make judgement that two things are alike or not we do so based on measurements we have performed. What measurements can we perform on these graphs?

Three concepts in particular stand out as robust measures of network topology [1].

1. *Average path length* is a measure of (on average) how many hops it takes to get from any chosen node to any other node.

2. The *clustering coefficient* is a measure of how connected a node's neighbors are to one another.[1]The measurements are typically averaged over all nodes in a graph.

*brainsik@cs.unm.edu, brainsik@theory.org

[1]It has been shown [1] that the clustering coefficient of a node can be reduced to the question: how many triangles is this node a member of. Since any single connection between a node's neighbors completes a triangle, this reduction makes sense.

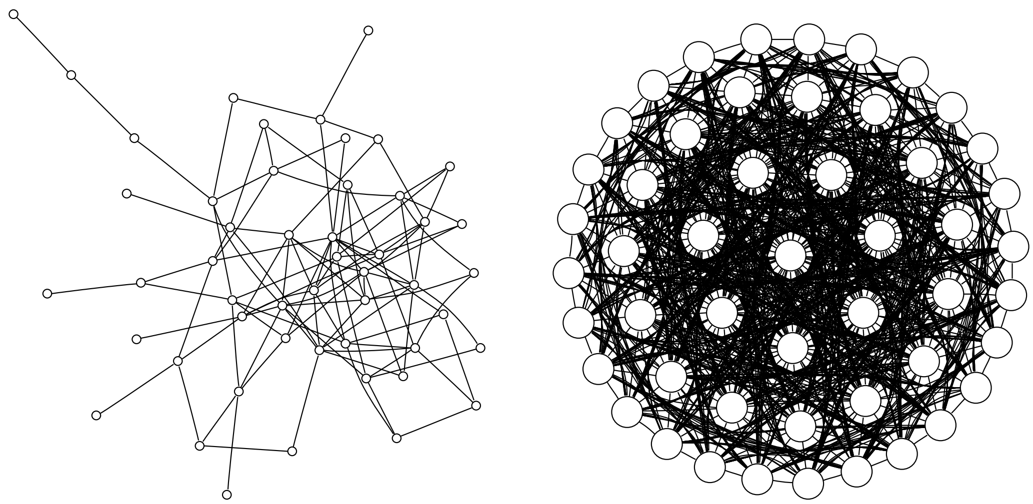

Figure 1: Classical Erdös-Renyi random graphs. Each graph contains 50 nodes. For the left graph, $p = 0.1$ and for the right graph $p = 1.0$ (producing a completely connected graph).

3. The *degree distribution* is a distribution function $P(k)$ giving the probability of a randomly selected node having exactly k edges (degree k).

A fourth measurement applicable to some networks is the size of the largest connected component, the *giant component*. Measuring the giant component can give much insight into how changing the parameters used to generate a graph can affect the topological properties of that graph.

2 Social Network Models

Four types of social network models were investigated with varying parameters: the classical Erdös-Renyi random graph [1], the Watts-Strogatz small world network [5, 6], a growing network by Callaway, Hopcroft, Kleinberg, Newman, and Strogatz (which I'll refer to as CHKNS) [2], and a growing network by David Lancaster [3].

2.1 Random Graphs

First presented in the 1950s, the traditional random graph (see Figure 1) is attributed to Erdös and Renyi. The random graph is generated using a value p which dictates the

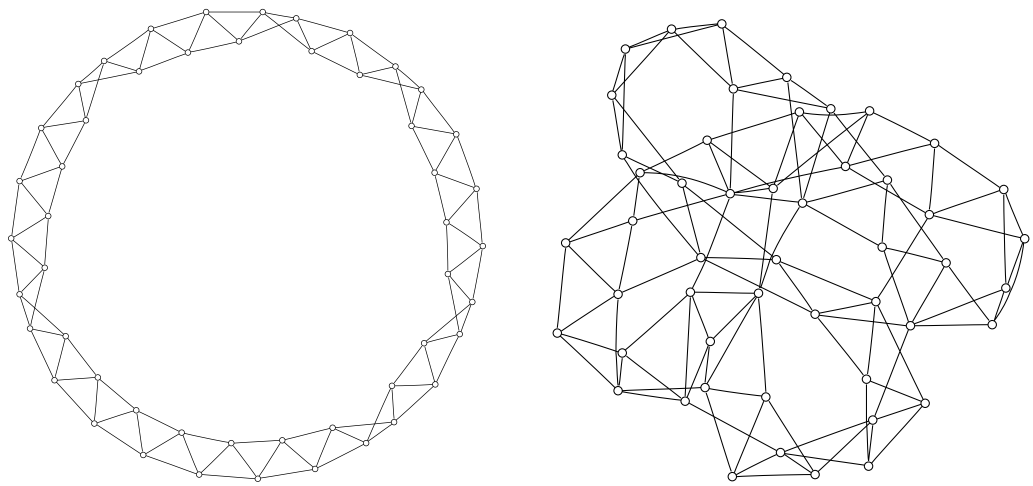

Figure 2: Watts-Strogatz small world graphs ($k = 2$). Each graph contains 50 nodes. For the left graph, $p = 0.0$ giving an ordered lattice and for the right the graph $p = 0.01$.

probability that any give node is connected to any other node. For $p = 1.0$, the graph is completely connected; every node is connected to every other node. Though the random graph captures the characteristics of very few real world social networks, it serves as a common benchmark against which modern models can be compared against.

2.2 Small-World Networks

Small world networks (see Figure 2) can be traced back to Stanley Milgram's original experiment (which initiated the phrase "six-degrees of separation") performed in 1967. In 1998, Duncan Watts and Steven Strogatz formalized the notions of small world networks [6] by presenting a tunable parameter p which could produce networks ranging from the ordered lattice to the random graph.

Small world networks are generated by starting with an initially connected lattice. The lattice has order k which indicates how many nearest neighbors each node is connected to. For $k = 1$, a simple ring is formed, each node connected to its nearest neighbors. Figure 2 shows the lattice formed when $k = 2$, each node connected to its second nearest neighbors. In the next step, this graph is treated as a classical random graph using the tunable parameter p (probability of linking). Having random links replace existing links or random links added to the set produces the same graph behavior.

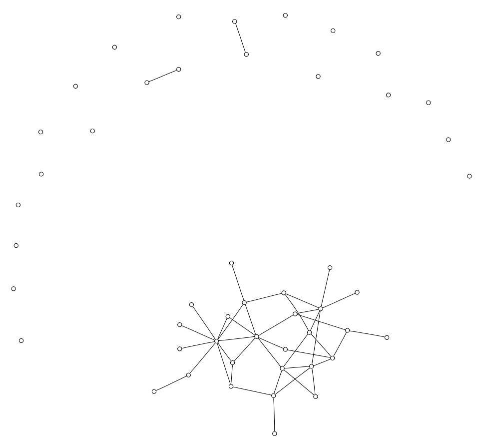

Figure 3: The CHKNS growing network model with 50 nodes and $p = 1.0$.

2.3 Growing Networks

Growing networks are the most recently introduced type of social network model. The primary difference between these models and the models described above is instead of starting with a fixed number of nodes to be connected together, new nodes are added over the lifetime of the network. The purpose of these models is to more closely simulate the formation of real world social networks (e.g., the WWW or telephone networks).

The original growing network models all use some form of preferential attachment [1] criteria. This criteria says that the likelihood a given node will receive a new edge is proportional to the number of edges it already has. However, the two growing networks investigated here are of a newer variety that try to capture the important phenomenon found in the original growing networks while remaining simple and avoiding the use of preferential attachment criteria.

2.3.1 The CHKNS model.

In 2001, a SFI working paper by D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz introduced a minimal model of a growing network [2]. For each step of the model, a new node is added to the network. Then, with probability p, two existing nodes are chosen uniformly at random and linked. Using this model, it was shown that randomly grown graphs are fundamentally different from statically generated random graphs. An example of a network generated with this model can be seen in Figure 3.

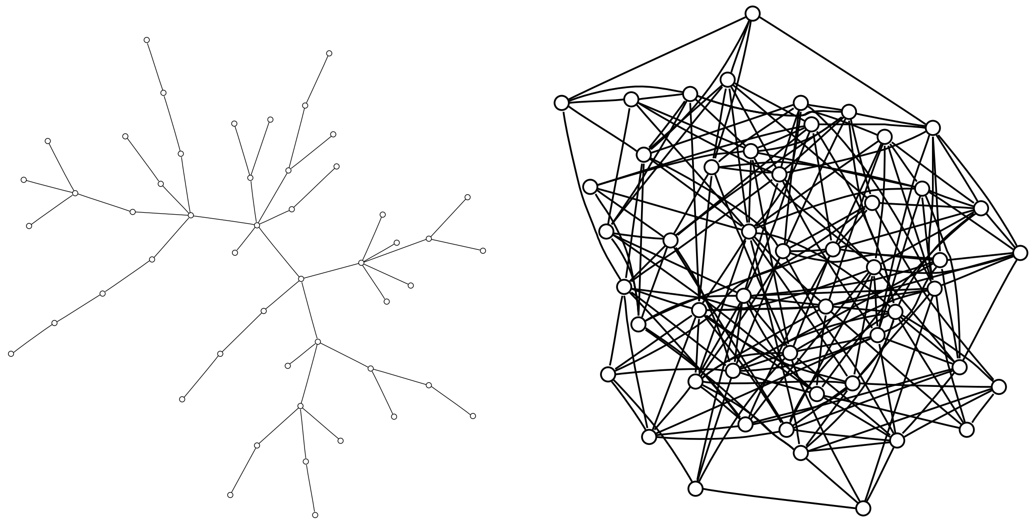

Figure 4: David Lancaster's growing network model. Each network contains 50 nodes and was generated with $p = 1.0$. The left graph used $k = 1$, the right graph $k = 2$.

2.3.2 The Lancaster model.

In 2002, a paper published by David Lancaster introduced a simple model for growing tree networks [3]. For each step of the model, a new node is added to the network. With probability p, this new node is linked to k nodes chosen uniformly at random from the pre-existing set of network nodes. Two example graphs generated using this model can be seen in Figure 4.

3 Interesting Discoveries

The large number of combinations of network and measurement types makes reporting all of them unreasonable. Following are some of the more interesting results I found while wandering the realm of social networks.

3.1 Growth of the Giant Component

Looking at the traditional random graph (See Figure 5), we see that there is a sharp phase transition where the network changes from being highly disconnected to being a single coherent piece. The sharpness of this transition is directly proportional to the number of

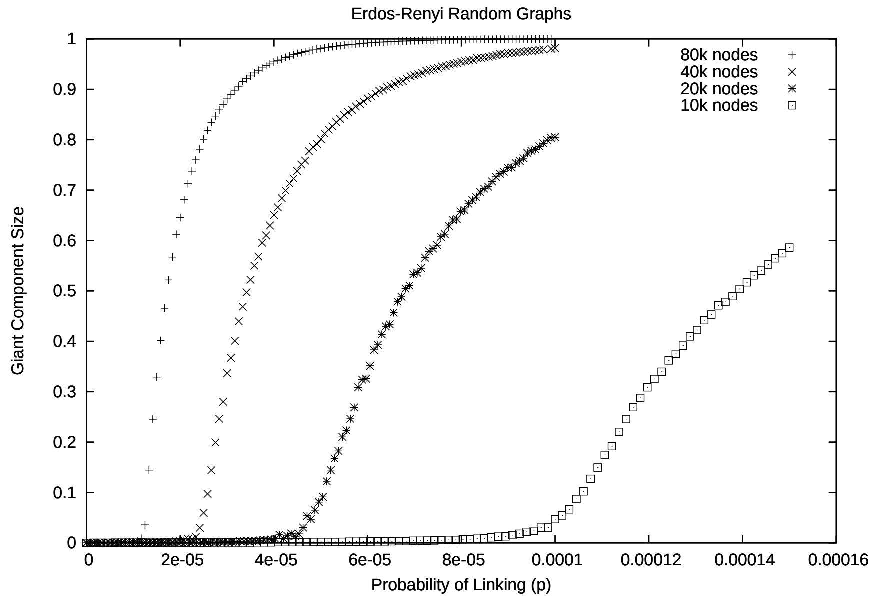

Figure 5: Every node has a chance of linking to every other node, once past a critical probability the giant component rapidly takes over Erdös-Renyi random graphs.

nodes present in the network. In particular, I found that as graph size doubles, the critical threshold halves. Only for very low values of p are random graphs not completely connected.

In contrast, although our two growing networks also show phase transitions for which the network changes from being highly disconnected to forming a large cluster, the size of this cluster ranges over the whole propability (p) spectrum (see Figure 6). Interestingly, the CHKNS model never forms a completely connected graph but instead has it's largest component occupying only 60% of the nodes. This is due to there being no gurantee, regardless of the probability of linking, that any node is linked to any other node. In the Lancaster model, when $p = 1$, every node has at least k links.

Suspiciously missing from the previous results above is the Lancaster model when $k = 1$. The way in which the giant component takes over this network with respect to changing probability is vastly different from the CHKNS model or for Lancaster models with $k > 1$. As can be seen in Figure 7, the network remains relatively disconnected until the probability of linking moves beyond 80%. Investigating the behavior of the giant component before it takes over the network reveals that it follows a power law related to the size of the network as it grows.

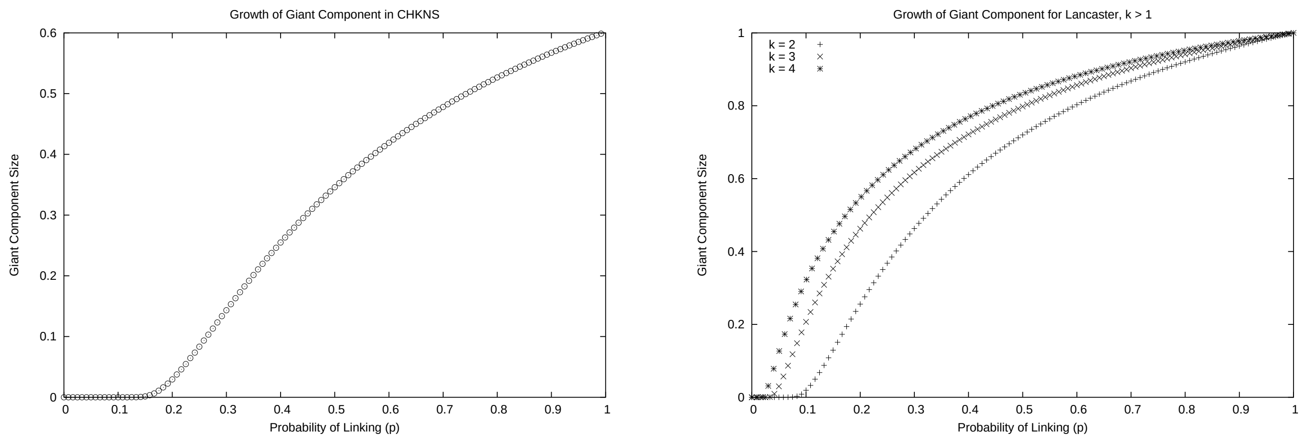

Figure 6: Growth of the giant component for the CHKNS and Lancaster growing network models. All graphs used 10 million nodes.

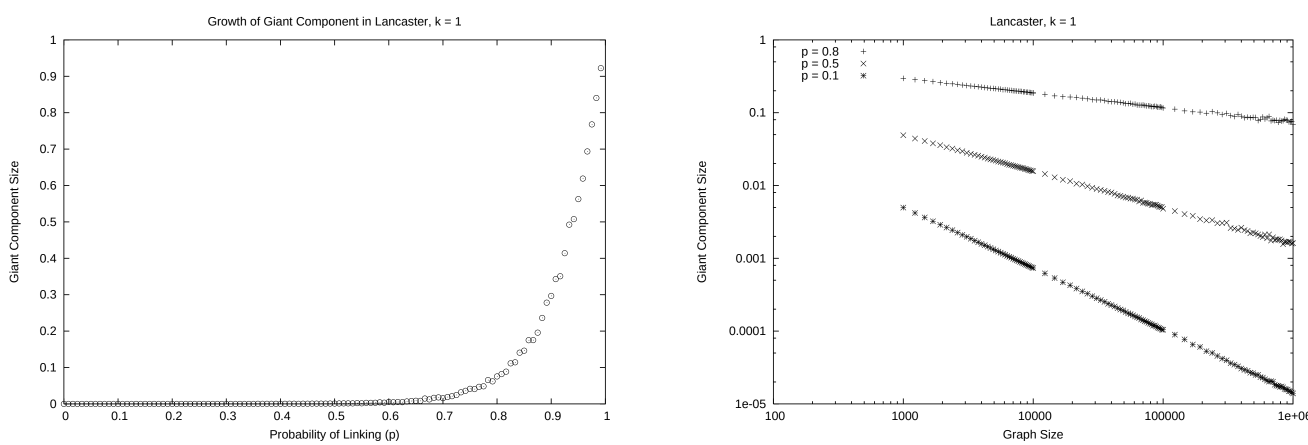

Figure 7: Before the giant component takes over the network, the Lancaster model exhibits a power law distribution in the size of the giant component as the network grows. Shown are the growths of the three giant components (right plot) for $k = 1$ type graphs where $p = 0.8$, $p = 0.5$, and $p = 0.1$. The critical exponents are $\gamma = 10.1$, $\gamma = 4.21$, and $\gamma = 2.48$ respectively.

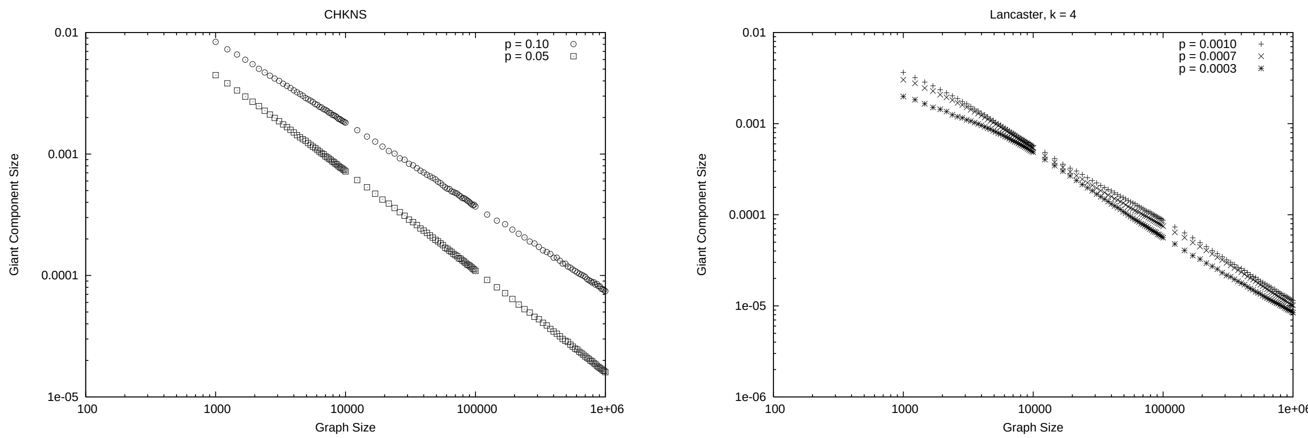

Figure 8: Before the giant component takes over the network, both the CHKNS and Lancaster $k > 2$ models display power law distributions in the size of the giant component as the network grows. For the CHKNS model on the left, $p = 0.1$ has $\gamma \approx 1.7$ and $p = 0.05$ has $\gamma \approx 1.35$. For the Lancaster model with $k = 4$, $\gamma \approx 1.7$ for all values of p.

The power law found in the Lancaster $k = 1$ model led me to investigate the other growing network models for p values below their giant component phase-transition. There too I found power law distributions for the size of the giant component versus size of the network (see Figure 8). For the Lancaster $k = 4$ model, I also found that varying p did not have a significant effect on the critical exponent.

The Watts-Strogatz small world models were not analyzed since, by defintion of their construction, they begin with a completely connected lattice. In otherwords, they begin with a giant component that occupies the entire network and since edges are not removed, this giant component will never change size.

3.2 Degree Distribution

Looking at the traditional random graph we find a Poisson distribution is formed (see Figure 9) soon after the giant component takes over the network (i.e. graph becomes fully connected). The Watts-Strogatz model looks similar since it starts with all of its nodes having the same degree and is then morphed into a random graph as its parameter p goes to 1.

In contrast, the two growing models show exponentially decaying degree distributions

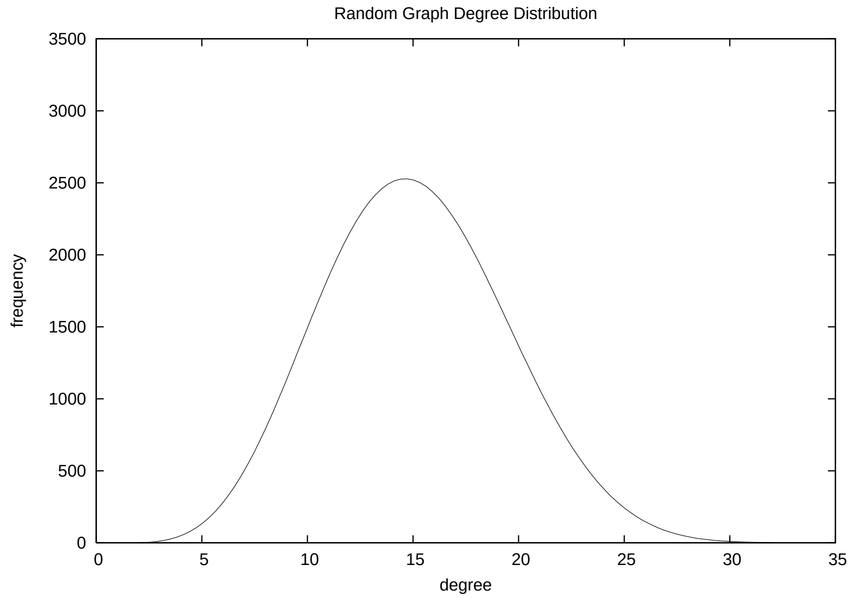

Figure 9: The traditional random graph displays a Poisson degree distribution right after the network becomes completely connected.

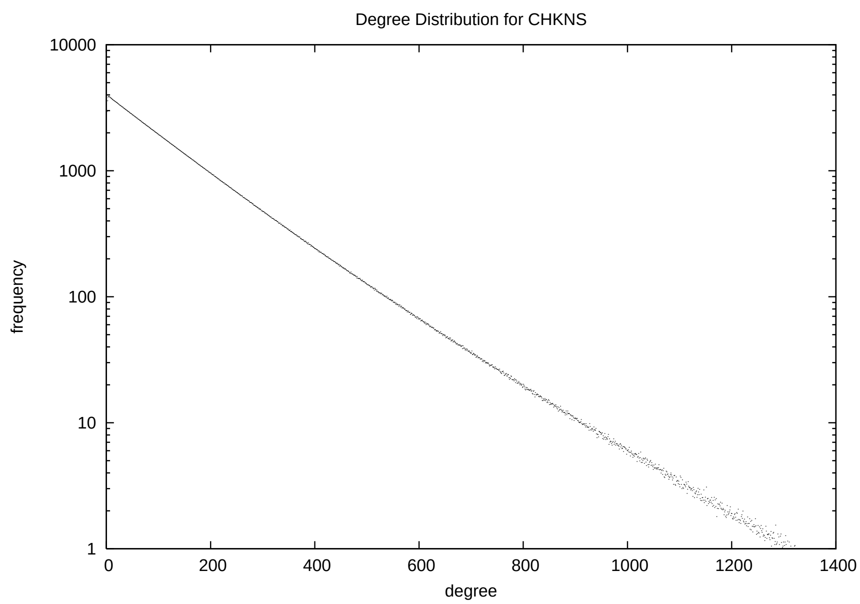

Figure 10: The CHKNS model has a degree distribution that decays exponentially with increasing degree.

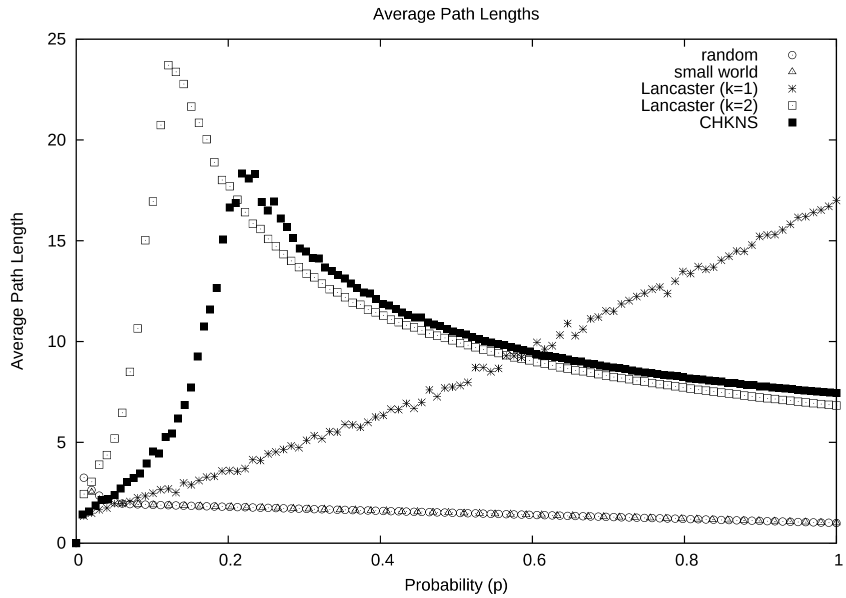

Figure 11: Average path lengths for the various social network models with a varying probability of linking. All path lengths are calculated for the graph's giant component.

(see Figure 10). Particularly interesting, the Lancaster models display a fixed frequency for degrees 0 to some threshold value depending on k.

The growing models are called *scale-free* [1] because they exhibit the same degree distribution regardless of network size. Scale-free networks were first described with networks grown with some preferential attachment critera. Those networks display a power-law in their degree distribution that is independent of network size.

3.3 Average Path Length

A composite plot (see Figure 11) reveals how average path length of the various social networks behaves with respect to the probability (p) of linking. This composite plot also reveals which network models behave similarly. We can see that the CHKNS and Lancaster $k = 2$ models have path lengths that steeply increase until they reach the cricital threshold seen for the growth of their giant components. The small world and the random graph follow identical, overlapping curves that live in the bottom of the plot. Lastly, the Lancaster $k = 1$ model is very different, showing a linear increase in path length.

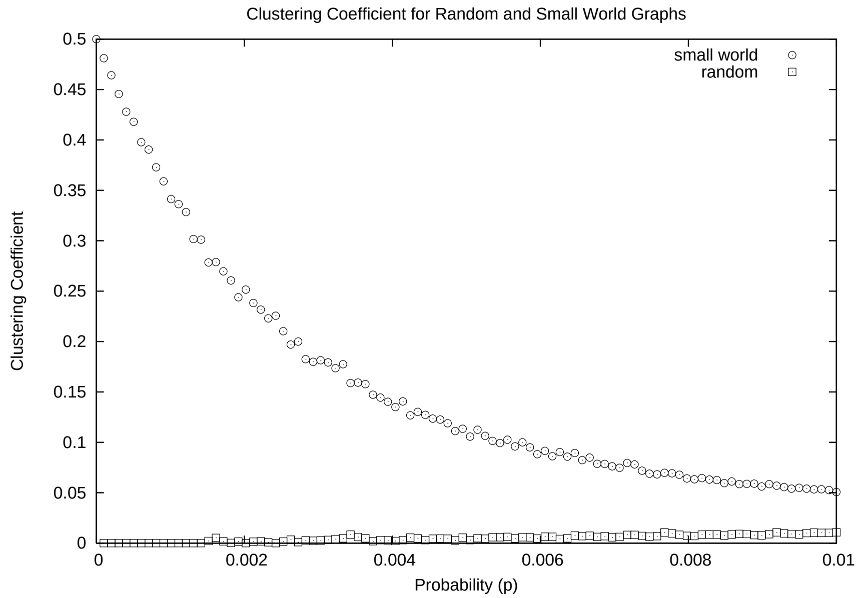

Figure 12: Small world graphs have much higher clustering coefficients than random graphs for small values of p. For larger p, enough random connections exist in the small world graphs that the clustering coefficient converges to that of the random graph.

3.4 Clustering Coefficient

In the previous section we saw that random graphs and small world graphs have almost identical path lengths. The quality that separates a small world graph from a random graph is the clustering coefficient. For small values of p, small world graphs have much higher clustering coefficients than random graphs. As p increases, the number of random links found in the small world graph increases, causing its properties to converge on that of a random graph.

Appendix: The Software

Having the boring moniker **rg** (originally meaning "random graph"), the software is able to statistically measure any one of a number of available social network models. The source code was written in C for speed and designed for easy addition of a new network models and statistical measurements. All of the software features are available from the command-line (i.e., one does not need to edit the source in order to access all of its functions), but the large

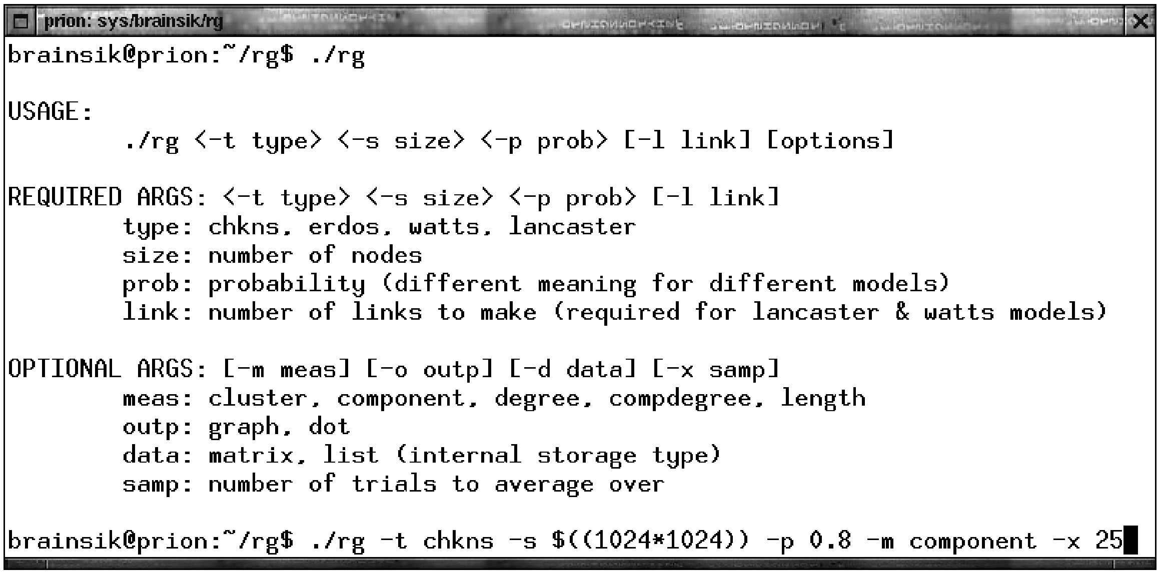

Figure 13: Screenshot of help screen and example command line.

space of options created by multiple network models, measurements and output methods can create an arcane command line (as seen in the figure). A simple (yet powerful) GUI could easily be written to interact via a command shell.

The software is able to generate four different network models.

1. Classical Erdös-Renyi random network (-t erdos) [1]
2. Callaway, Hopcroft, Kleinberg, Newman, and Strogatz growing network (-t chkns) [2]
3. Watts-Strogatz small world (-t watts) [5]
4. David Lancaster's growing network (-t lancaster) [3]

Each model can have a variety of different characteristics set.

1. Size of the network; the number of nodes (-s)
2. Probability of linking (-p)
3. Number of links to make (relevant to Watts-Strogatz and Lancaster models) (-k)

On each model, up to five statistical measurements can be calculated [1].

1. Clustering coefficient (-m cluster) [5]
2. Size of the giant component[2] (-m component)
3. Whole graph degree distribution (-m degree)
4. Giant component degree distribution (-m compdegree)
5. Average path length[3] (-m length)

The time and space requirements as well as the accuracy of the measurements can be affected. The time and space requirements can be modified for some of the measurements by explicitly setting the type of internal graph representation used: adjacency matrix (-d matrix) or adjacency list (-d list).[4]The accuracy of the measurements can be improved by averaging the values calculated over a series of trials (-x).

Output can include two different graph representations in conjuction with the results of the measurements (which are always displayed). The first type of graph representation is ascii view of the graph data in its adjacency matrix form, adjacency list form, or both. The second type of graph representation is in GraphViz[5] dot notation for use with the GraphViz open source graph drawing software.

In addition to this core network generating and measuring software is a Perl script, **wrap.pl**. This wrapper is used to easily genereate and measure a network with one of its characteristics (such as size or probability) being varied over a specified interval for specified number of steps.

All source code is freely available under the GNU Public License[6] (GPL) and will be made available via my home on the UNM CS webserver[7] as well as Theory.org[8] .

[2]Calculated using the very fast Union-Find algorithm.
[3]Calculated using my own variant of BFS (breadth-first-search).
[4]Some measurements require specific data structures to be used in their calculation.
[5]Graphviz - open source graph drawing software. `http://www.research.att.com/sw/tools/graphviz/`
[6]GNU General Public License. `http://www.gnu.org/licenses/gpl.html`
[7]Brainsik at UNM. `http://www.cs.unm.edu/~brainsik/`
[8]Theory.org homepage. `http://www.theory.org/`

References

[1] Albert, R. and Barabási, A.-L., 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47-97.

[2] Callaway, D. S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. J., and Strogatz, S. H., 2001. Are Randomly Grown Graphs Really Random? SFI Working Paper, 01-05-025.

[3] Lancaster, D., 2002. Cluster growth in two growing network models. *J. Phys. A* **35**, 1179-1194.

[4] Strogatz, S. H., 2001. Exploring complex networks. *Nature* **410**, 268-276.

[5] Watts, D. J., 1999. *Small Worlds.* Princeton University, Princeton.

[6] Watts, D. J. and Strogatz, S. H., 1998. *Nature* **393**, 440.