Network Analysis and Modeling, CSCI 5352 **Prof. Aaron Clauset**
Lecture 8 **24 September 2013**

1 Comparing networks, and why social networks are different

The various measures of network structure that we have encountered so far allow us mainly to understand the structure of a single particular network. In many cases, however, we wish to compare the structure of two or more networks. For example, are some pair of social (or biological) networks similar? Are there general patterns in network structure that distinguish different entire classes or groups of networks? Are social networks different from non-social networks?

The question of how to compare networks remains an important and unresolved task in network science. That being said, a common approach is to compare networks across a consistent set of scalar measures, such as those we have seen so far. This approach effectively takes each element of a set of graphs $\mathcal{G} = \{G_1, G_2, \ldots, G_r\}$, each of which is a complicated, non-Euclidean mathematical object, and projects them into some vector space. Given r vectors in the space defined by the network measures, we may apply conventional measures of similarity and distance. A particular graph's coordinates within this space is determined by the values of its network measures.

Crucially, not all such vector spaces are equal. Many network measures are correlated, as we saw with centrality measures. The implication is that correlation is not sufficient to identify interesting underlying properties of some class of networks or to identify general differences among classes of networks.

The Table on the next page[1] lists a number of different empirical networks, drawn from social, informational, technological, and biological systems, and their values along a number of network measures. The field has progressed greatly since this paper was published in 2003, but there are relatively few other or newer "big tables" showing a specific set of network statistics for a large and diverse set of real-world networks. As such, it remains an instructive list with many lessons to teach.

In particular, notice the differences between social and non-social networks in both the clustering coefficients[2] and the degree correlation coefficient (degree assortativity).

In particular, the clustering coefficient to away from zero in social networks, but very close to zero in non-social networks, and similarly the degree correlation coefficient tends to be positive in social networks, and negative in non-social networks. This is indeed, one of the main differences between social and biological networks: in social networks, we observe that vertices with similar degrees tend to be connected, while in biological networks, vertices of unlike degrees tend to be connected.

[1] Reprinted from M.E.J. Newman, "The structure and function of complex networks." *SIAM Review* **45**, 167–256 (2003).

[2] The first clustering coefficient $C^{(1)}$ represents the version given in a previous lecture: the fraction of connected triples that are themselves triangles. The second clustering coefficient $C^{(2)}$ is the mean local clustering coefficient, i.e., $C^{(2)} = n^{-1} \sum_i C_i$, which is not a common definition of clustering, but is sometimes used.

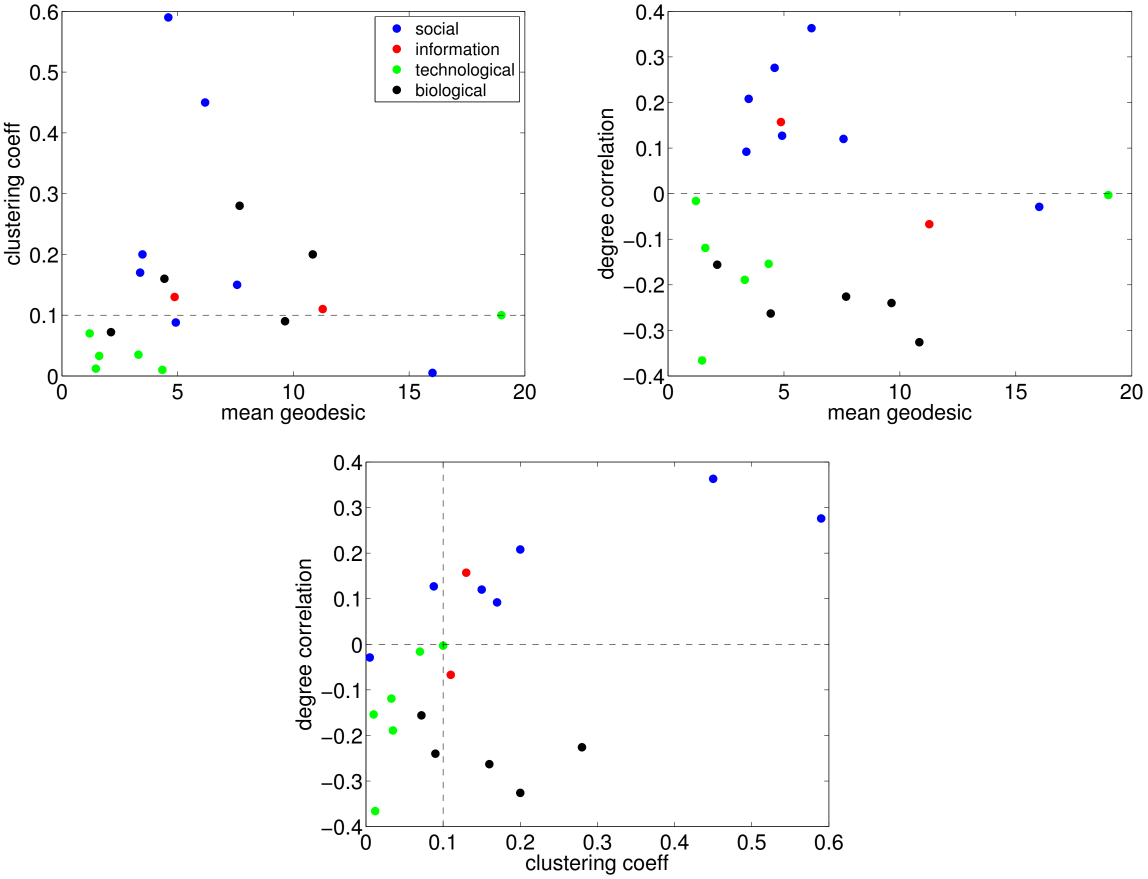
	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

And, we observe highly non-trivial amounts of localized transitive closure in social networks, in which friends of mine also tend to also be friends with each other.

To visualize these correlations and patterns more clearly, the figures on the next page show the pairwise scatter plots of the mean geodesic distance ℓ, clustering coefficient $C^{(1)}$, and degree correlation coefficient r for networks in this table. (Networks missing data on any one of these three are omitted.) Important thresholds, such as a $C^{(1)} < 0.1$ or $r = 0$ are shown as dashed lines. Most notably, social networks show positive degree correlations, while technological and biological networks show negative correlations; social networks show the highest clustering coefficients (but also some small ones), while technological networks have small coefficients.

There is generally less agreement or difference along other network measures (including those shown in the table, but also on a wide variety of other measures). For instance, networks of all different types exhibit small or large mean geodesic distances, and tend to come in a wide variety of sizes. Similarly, while some networks exhibit power-law degree distributions, many seem to exhibit non-

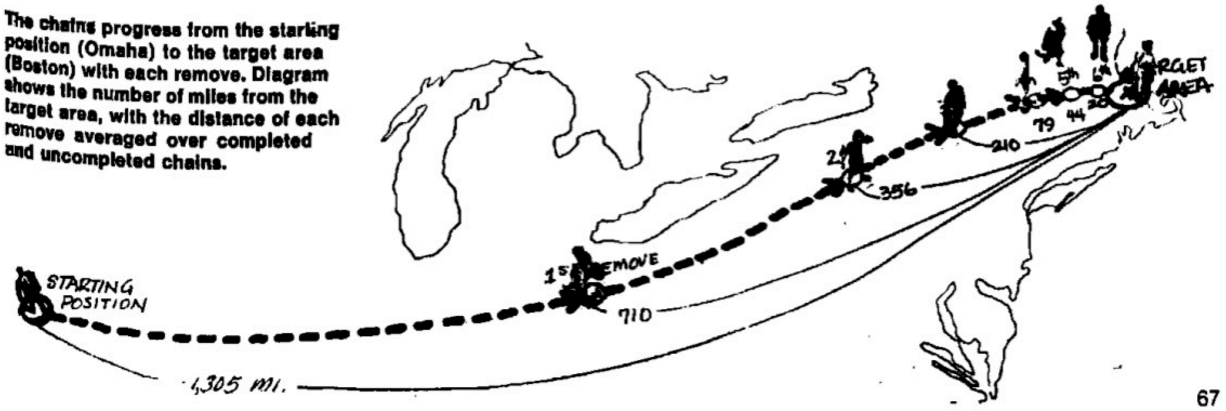


power-law distributions. (However, nearly *all* networks exhibit heavy-tailed degree distributions.) There has not been a recent systematic survey of network measures with respect to classes of networks, but these above insights seem to have been borne out by more recent analyses.

2 Small worlds, navigability, and social networks

An important, and arguably under appreciated property that distinguishes social networks from other types of networks is their "navigability," a property that is related to a network's diameter and the "small world" property.

The idea of a small-world network comes from a seminal study in social networks by the American



The chains progress from the starting position (Omaha) to the target area (Boston) with each remove. Diagram shows the number of miles from the target area, with the distance of each remove averaged over completed and uncompleted chains.

sociologist Stanley Milgram (1933–1984).[3] Milgram mailed letters to "randomly selected" individuals in Omaha, Nebraska and Wichita, Kansas with instructions asking them to please pass the letter (and instructions) to a close friend of theirs who either knew or might be likely to know a particular doctor in Boston (see figure above).[4] Before doing so, they should also write their name on a roster to record the chain of message passing. Of the 64 letters that eventually reached the doctor—a small fraction of those sent out—the average length was only 5.5, and a legend was born.

There are two interesting things about Milgram's study.[5] First, if we näively expected a "big world" social network, then the number of steps between the source and the target should be something much larger than 5.5. Imagine that people were only friends with individuals a few miles down their road. In this case, the social network's structure would be similar to that of a 2d lattice. The path length between Omaha/Wichita and Boston would then be proportional to the number of people living between those points on the map, and which would yield a length of hundreds or more individuals, not a handful. This idea, that the network has small average pairwise distance, is the informal notion of a "small world," which is usually formalized mathematically as a diameter that grows like $O(\log n)$ with the size of the network.

Second, despite each individual in the chain having only local information about their own neighbors in the social network, some letters somehow managed to arrive at the destination. The letters that

[3] The term "six degree of separation" is not due to Milgram, but comes from a play written by John Guare in 1990. The play was subsequently made into a movie of the same name starring Will Smith in 1993, and, ironically, not starring Kevin Bacon.

[4] Figure reprinted from S. Milgram, "The Small-World Problem." *Psychology Today* **1**(1), 61–67 (1967).

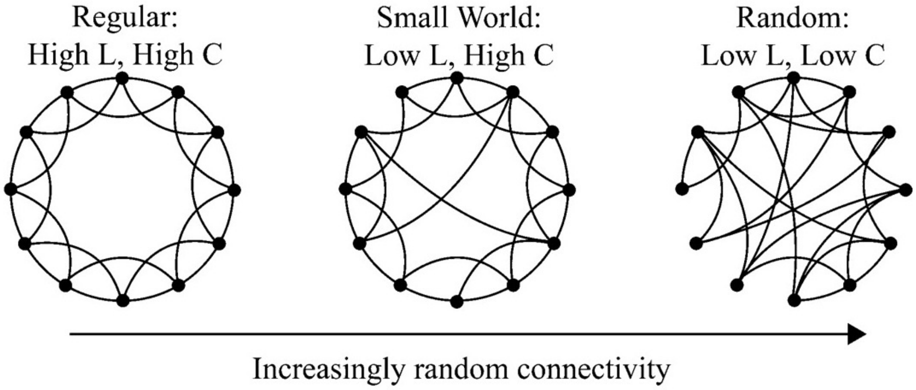
[5] Milgram is famous for another study, as well, on obedience and authority. For a shocking overview of his results, see this `http://www.youtube.com/watch?v=Jqr5-dWk6Gw` .

failed to arrive could not find a path from the source to the target, because some individual on the path failed to forward the letter (perhaps because they did not know anyone "closer" to the target or because they weren't interested in participating in the study). Thus, the arrival of a non-trivial number of letters demonstrates that not only do short paths exist (small diameter), but these paths can also be found by individuals using only local network information. We call this property called *navigability*, and we will return to it in the next lecture.

2.1 The Watts-Strogatz model

The first of these properties, the existence of short paths, was not all that surprising by itself. In fact, the classic model of random graphs, which we will learn more about starting next week, exhibits the small world property and has a diameter of $O(\log n)$. However, random graphs exhibit almost no local clustering. In contrast, lattices exhibit high local clustering, but have large diameters. How can a graph have both small diameter *and* high clustering?

In an early paper in network science, Duncan Watts and Steve Strogatz studied this specific question using a simply toy model[6] that interpolates between these two types of networks. This model is sometimes called the "small world" model.



In this model, n vertices are arranged on a 1-dimensional circular lattice (a "ring" network) and each vertex is connected to its k nearest neighbors. The left-most network in the figure above[7] shows such a lattice with $k = 4$. Given this starting point, for each lattice edge (i, j), we then

[6] A "toy" model is a simple mathematical construct that is useful for building intuition and studying specific phenomena. It is not meant to be a realistic model of real-world networks. Toy models are popular in Physics.

[7] This figure, and the one on the next page, both reprinted from Watts and Strogatz, "Collective dynamics of 'small-world' networks." *Nature* **393**, 440–442 (1998).

"rewire" it uniformly at random with probability p. (By rewire, we mean change j to k, where k is chosen uniformly at random from among all vertices.) As $p \to 1$, all edges are rewired, and we have a classic random graph, with no local clustering but small diameter. When $p \to 0$, we have the original lattice, with high local clustering but large diameter.

The interesting behavior emerges as we move p between these two limiting cases (see the figure below). When only a small fraction (p) of edges have been randomly rewired, the diameter of the network collapses from $O(n)$ to $O(\log n)$. This happens, however, long before the local clustering decreases. Thus, the intermediate, "partially disordered" states are those in which the network exhibits high local clustering and short paths between vertices. This is exactly the counter-intuitive behavior of social networks, which exhibit very strong local clustering at the same time as, from Milgram's study, showing short paths.

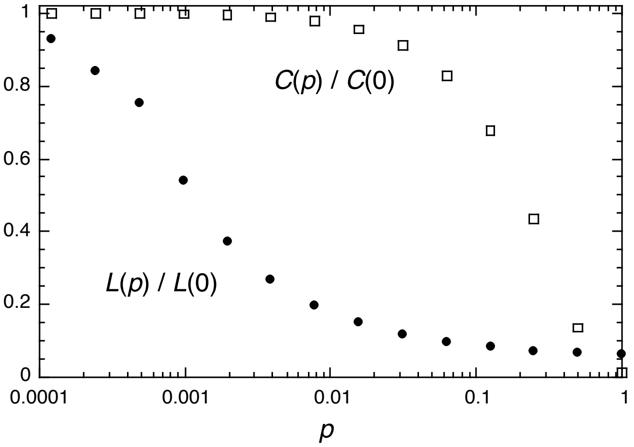


Figure 1: As a function of the rewiring rate p, the clustering coefficient C (normalized by the original clustering coefficient $C(0)$) and the mean geodesic path length L (also normalized by the original mean geodesic path length $L(0)$).

In addition to being an early example of network science, this result is interesting for another reason. In particular, it shows that some measures of network structure are extremely sensitive to small variations in network structure. For the small-world toy model, with less than 1% of the original edges have been randomly rewired, the diameter has already fallen to roughly 20% of its original value, while the clustering coefficient has barely budged. Thus, path lengths are extremely sensitive to the presence (or absence) of some edges, while clustering coefficients are highly robust to such variation. Thus, if an empirical network is sampled, i.e., each node and/or edge is observed with some probability, we may have a dramatically incorrect view of the true network structure.

3 At home

1. Reread Chapter 8.2 (pages 241–242)

2. Read Chapter 15.1 (pages 552–564) in *Networks*

3. Next time: navigable networks