An overview of the study of networks

May 3, 2005

The paper first describes a variety of network types, then lists common network metrics, and finally how they interrelate.

It is based heavily on Small Worlds by Duncan J Watts; page numbers refer there.

1 The basics

Notworks consist of individuals connected by links. Links are from exactly one person to exactly one other, are bidirectional, and are all of equal strength. That is, the link between two nodes is binary: either it's a two-way full strength link, or is entirely missing. Networks with partial links are not as well-studied.

Finally, throughout we will want the network to be connected, meaning that any node can somehow be reached from any other node. Information obviously can not travel between two disconnected segments of a graph, but that's boring.

2 Types of artificial network

Physical networks The easiest kind: link people who are close together in a geographic space. Space may be one-, two-, or n-dimensional, and may be flat or a circle/torus. Links may be deterministic (everyone within s steps of you is linked) or probabilistic (probability of being linked decreases with distance).

Pros/cons: works great iff people really are actually spatially distributed. Does a bad job describing things like the Bacon network.

Random graphs Watts offers two types (p 36):

$G(n, M)$ refers to a graph with n nodes and M edges chosen entirely at random. $G(n, p)$ refers to a graph where each edge exists with probability p; that is, we do an independent draw for each edge to determine whether it will exist.

Pros/cons: easy baseline, but no *ex ante* reason to think that links randomly form in any real-world situations; notably, we will see that clustering and popular nodes are much more common in reality than in a random graph.

Grow the network A Waring network is constructed by adding links one by one. Begin with a set of nodes with no links, add links by the following two rules:
For each node in turn,{

- Connect to someone with zero links with probability $(\beta - 2)/(\beta + \alpha - 1)$.
- Assuming the new link was not to a zero-link node, then the probability that it will be to a link with η existing nodes is $\propto \eta f(\eta)$, where $f(\eta)$ is the percentage of nodes with

exactly η links. We need to recalculate $f(\eta)$ every iteration.
} until the link density or link count is as desired.

Notice that a node with a large number of links is more likely to get linked to. In colloquial parlance: the rich get richer.

Pros/cons: Tells a reasonable story about how networks grow from nothing. Link density has a very specific form.

The α-model A key player in Watt's story, pp 46-47.
First, fix the average degree that the network will have to k. Let p be a very small, fixed value. Then:

```
While (link density so far $< k$)
    for each node $i$
        for each node $j \neq i${
            $m_{i,j}$ = nodes linked to both $i$ and $j$
            if ($m_{i,j} \geq k$) link with certainty
            if ($k > m_{i,j} > 0$) link with probability $[m_{i,j}/k]^\alpha(1-p)+p$.
            if ($m_{i,j} = 0$) link with probability $p$.
        }
```

This is parametrized by α: for $\alpha = 0$, you get the 'caveman' network, with small connected graphs which have no connections of any sort; for moderate α, the number of connections between caves grows, and the caves themselves are less likely to be complete graphs; for $\alpha \to \infty$, the caves are erased entirely, and people are as likely to link to a friend of a friend as to a complete stranger—the network approaches a random graph.

Guaranteeing that the network remains connected is a problem, especially for small α; Watts suggests beginning with a substrate of some sort, such as a ring, and then adding enough links that we can hope that the substrate becomes irrelevant. On pages 48–, he attempts to convince us that the substrate will be irrelevant for the analytics.

Pros/cons: it's parametrized, so we can ask how network metrics vary as a function of α. For large values of α, we can compare to a random network. Sort of tells a story of how the network evolves.

The β-model Instead of growing the network, begin with a substrate and then randomly rewire it. Begin with a 1-lattice: everyone is in a ring, connected to their next-door neighbors and their next-to-next-door neighbors. (four links total).

```
for $c$ in 1 to $n/2$
    for each node $i$
        if $i$ is linked to person $i+c$
            with probability $\beta$ delete the link and randomly assign it
            to anyone else (who isn't already linked to $i$).
```

If $\beta = 0$, nothing every gets rewired and we stay with our lattice; if $\beta = 1$, every link encountered gets rewired every time, and we approach a random graph. The book cover shows three graphs beginning with $\beta = 0$.

Pros/cons: it is parameterized, with random graphs at one end of the parameter space, but there is no realistic story about how such networks would form.

3 Network metrics

link density The easiest measure of link density is just k: the number of links divided by the number of nodes.

path lengths Let $d(i, j)$ be the shortest distance between i and j, and $\bar{d}_i$ = the mean of $d(i, j)$ over all $j \neq i$. Then the *characteristic path length* $L(G)$ is the median of $d(i, j)$. We want the median because links are typically very asymetrically distributed, see below.

The *diameter* of a graph is the maximum value of $d(i, j)$ in the graph, i.e., the longest shortest path. That is, if an infection or idea appears somewhere in the graph, what is the longest it would take to cover the graph?

For most applications, we care about the path length metrics more than any others. Public health studies want to know the speed of disease transmission, which is most heavily influenced by k and $L(G)$. Advertisers, revolutionaries, and die-hard democrats want to know how long it would take for a fact to spread by word-of-mouth.

Clustering How cliqueish is the graph? Let $\Gamma(i)$ be the nodes immediately connected to i—its neighborhood, and Let γ_i be the number of links in $\Gamma(i)$ divided by the possible number of links, $\binom{n}{2}$, where n is the number of nodes in $\Gamma(i)$. Then the *clustering coefficient* for the whole graph, γ, is the mean of γ_i over all i.

Connectedness The basic idea to all of these measures is that if there were a bit of network damage—a few nodes or links die—how much longer would path lengths be?

If the direct link between i and j were missing, then if $d(i, j) > 2$, then the link taken to be missing is called a *shortcut.* ϕ is the percentage of links which are shortcuts.

ψ is the percentage of nodes which are not connected but which share one and only one neighbor. That is, $d(i, j)$ is two, but if a certain node is deleted, $d(i, j) > 2$.

We can define the characteristic path length of a set of nodes as we did above: what is the median of the mean of $d(i, j)$ for everyone in the neighborhood. The connection between the two nodes may involve links that leave the neighborhood and then re-enter. Then the *significance* of a node, $S(i)$ is $L(\Gamma(i))$ if i were deleted. The significance for the whole graph G, $S(G)$, is the mean of $S(i)$ over all i.

Some of these measures are better than others for different situations. In the Bacon graph, everyone who worked on a movie is connected to everyone else—that is, each movie is defined to be a complete graph—meaning that $\phi \approx 0$, but ψ is still significantly different from zero.

These metrics are basically unrelated to path length, in the sense that one could draw up a graph with high L/low ϕ, low L/high phi, et cetera.

distribution of links [Watts doesn't mention this stuff.] Let $\rho(x)$ be the number of nodes with x links. For a ring, $\rho(2) = n$ and $\rho(x) = 0$ for all $x \neq 2$. For a star, $\rho(1) = n - 1$ and $\rho(n - 1) = 1$. For the random graphs above, $\rho(x) \propto$ binomial(n,x).

If $\rho(n) = k^{-n}$ (where k is an arbitrary constant), then we say that the links follow a *power law*, or are Zipf distributed.

The wonderful thing about Waring and Yule's network-growing methods is thaT they limit to a closed-form link density. Here's the link density for the Waring distribution:

$$\rho_w(n) = \frac{(\beta - 1)\Gamma(\beta + \alpha)}{\Gamma(\alpha + 1)} \cdot \frac{\Gamma(n + \alpha)}{\Gamma(n + \alpha + \beta)}.$$

The Yule distibution converges to the same with $\alpha = 0$:

$$\rho_y(n) = \frac{(\beta - 1)\Gamma(\beta)\Gamma(n)}{\Gamma(n + \beta)}.$$

$\rho_y(n) \approx \beta^{-n}$; i.e., these models approach a power law.

The Yule, Waring, and Zipf distributions look a lot alike: there are a lot of nodes with only one link, and very few nodes with many links. But the tail goes much further out than for the binomial distribution.

4 Selected interactions

Random graphs The characteristic path length L for a random graph is typically very small. That is, information tends to travel very fast in a random graph. Too bad people don't join up via a purely random mechanism.

There is no real clustering to speak of in a random graph beyond random aggregations; $\gamma = \frac{k-1}{n}$; since k is typically orders of magnitude smaller than n, so $\gamma \to 0$ for a random graph.

γ, L, and α Both $\gamma(\alpha)$ and $L(\alpha)$ have the same shape: they start high, and then quickly fall. But Watts's key observation is that $\gamma(\alpha)$ shows its sudden drop-off around $\alpha = 5$, while $L(\alpha)$ drops off around $\alpha = 8$. That is, for $\alpha \in (5, 8)$, the graph is much more clustered than a random graph, but the characteristic path length is very short. This is the *small world* graph which interests him (and which is formally defined as a graph with $L \approx$ a random graph but high γ on p 114).

Scaling What happens as n gets large? Since many of the above statistics require $n!/2$ computations, it would be nice to have an equation as a function of n to save us the trouble of calculating them.

For distributions built using the rich-get-richer algorithms, the small world graphs, and the random graphs, $L(G)$ and $D(G)$ effectively do not vary with n. This is where the six-degrees-of-separation stories come in: even with billions of people, $D(G) < 10$—the same order of magnitude as $D(G)$ for groups where n is a few dozen or hundred.

Compare with physical networks: if everyone lives strictly on a 2-d grid, $D(G)$ is proportional to the square root of the size of the grid; a hundred person network has $D(G) = 10$, and a one-million person network has $D(G) = 1,000$.

Rich-get-richer type graphs are typically called *scale-free networks*, but random graphs and the α-graph with $\alpha > 5$ also have L which effectively does not vary with scale, but are typically not called scale-free.

fragility It may be worth distinguishing between two types of network breakage. Natural breakage would be the death of links selected entirely at random. For a rich-get-richer type graph, natural breakage is no problem, because the majority of links have only a few links. Human breakage would be a concerted effort to break the system; for this, the rich-get-richer graph is supremely fragile, because breaking a few hubs is sufficient to significantly increase D.