

Lecture 23 Networks

July 28, 2006

1 Background

Even before the growth of the Internet and the associated focality of the “network” metaphor, network theory was a large and active research area within several disciplines most notably mathematics, ecology, and sociology. It would be impossible to do justice in one lecture to an idea that spans inter species consumption (food webs), transportation (traffic networks), power transmission (electric grids), job seeking (social networks), political beliefs (influence webs), group formation (cluster analysis), disease spreading (interaction structures), advertising (communication networks), and cognition (neural nets). Accordingly, we will restrict our focus to topic especially relevant to diversity. The three topics we will discuss are (i) diversity and connectedness (ii) multiple paths and robustness and (iii) strategic choice of diverse connections.

We begin with some of the basic terminology and ideas from social network theory. We use this rather than an alternative vocabulary such as from graph theory because social network theory relies on more familiar terminology.

2 A Social Network Theory Primer

As in our previous discussion of perspectives and interpretations, we begin with a set of objects.

Def'n: The set of **objects** equals $\Omega = \{a, b, \dots\}$

In graph theory, these objects would be referred to as *nodes*. The objects have *relations/connections* between one another. These relations can be in one direction: I know who Oprah Winfrey is, but she does not know me; Snakes eat mice, but mice do not eat snakes. Or they may be two directional: Churchill and Roosevelt were friends; Ford and General Motors compete for market share, etc. . . . In keeping with our goal of simplicity we will throughout this analysis that all of the connections are two directional.

Def'n: A **connection** χ is a pair of objects $\chi = (a, b)$ from Ω

Using this terminology, three people Alice (A), Babu (B), and Carl (C) who are all friends with one another can be described as three connections (A,B), (B,C), and (A,C). Similarly, if Babu is friends with both Alice and Carl, but Alice and Carl have never met, then the network can be described as two connections (A,B) and (B,C). Notice that in this second network, Babu's placement in the graph is more central.

We can use this notation to describe any two directional network.

Def'n: A **network** N is a set of objects Ω together with a set of connections χ , $N = (\Omega, \chi)$.

Now that we have defined networks, we can define the *neighborhood* of an object a within a network. These are the objects which are connected to a .

Def'n: The **neighborhood** of object a in the network N , $n(a)$, equals the set of all objects that are connected to a .

In our one dimensional spin glass model from the previous lecture, the neighborhood of an agent consisted of the agents to the agent's right and left. In social networks, your neighborhood is the people you know directly. We can extend the idea of neighborhoods to characterize the people who know people you know – friends of friends. We call this the *two-neighborhood*.

Def'n: The **two-neighborhood** of object a in the network N , $n^2(a)$, equals the set of all objects that are connected to objects in $n(a)$

This same idea can be extended indefinitely. A person's three neighborhood consists of friends of friends of friends. These secondary and tertiary friends prove empirically important in getting jobs, meeting soul mates, and finding information as was detailed in a now famous paper by Mark Granovetter called *The Strength of Weak Ties*. As we will see in a few minutes, as we move further out the network to 4-neighborhoods under rather mild assumptions on the topology of friendships, we quickly get huge sets of people. This fact underlies the famous six degrees of separation idea which we discuss momentarily. But first, we need more formal definitions.

Def'n: The **k-neighborhood** of object a in the network N , $n^k(a)$, equals the set of all objects that are connected to objects in $n^{k-1}(a)$

In light of this construction, we would be more precise if we referred to our *neighborhood* as our *one-neighborhood*, but doing so seems overly pedantic. Notice that the k -neighborhoods do not form a partition over the objects: someone who belongs to my neighborhood might also belong to my two-neighborhood and to my three-neighborhood. Therefore, we define the *distance* between two objects a and b as the smallest k such that a belongs to the k -neighborhood of b .

Def'n: The **distance** between object a and b in the network N , equals the smallest k such that b belongs to $n^k(a)$

Just knowing the distance between two objects may not be sufficient information. We might also want to know how to traverse the network to connect the objects. To capture this idea, network theorists developed the *minimal path* concept.

Def'n: A **minimal path** between object a and b that are separated by distance k is a sequence of objects $(a_0, a_1, a_2, \dots, a_k)$ where $a_0 = a$ and $a_k = b$ and a_i lies in $n(a_{i-1})$ for $i = 1$ to k .

Minimal paths may not be unique. If a is friends with b and c and if both b and c are friends with d but a is not, then a has two minimal paths to d , abd and acd . We will show that in some cases having multiple minimal paths may be a good thing.

So far, we have not said anything about these objects. If the objects are people, they possess characteristics: age, gender, ethnicity, etc.. These characteristics may partially determine the networks structure. In instances where the networks form by choice, the tendency for people to make connections with other people like them is well documented, so much so that sociologist Robert Merton coined the term *homophily* to describe it. If in contrast, networks are enforced organizationally, there could be an emphasis in the opposite direction - toward diverse connections.

3 Network Connectedness

The idea that most people in the world are connected through paths of length six or fewer gained popularity through the play *Six Degrees of Separation* and through the tongue in cheek application of the idea in the form of a game (and subsequent web page) called "Six Degrees of Kevin Bacon" in which participants must find minimal paths between the actor Kevin Bacon and other actors and actresses. The original experiments on degrees of separation were conducted by Stanley Milgrom who had people in a small Midwestern town attempt to get a letter to a banker in Boston through their social networks. The typical letter's path required about six postings.

We begin by relating the six degrees finding to the structure of networks. We begin by making some stylized assumptions. First, we'll assume that each person has

exactly 150 friends. This will make the mathematical calculations easier. It's also close to accurate. The average number of friends does average around 150 (Ridley 1998). We will also assume a total population of six billion, a six followed by nine zeroes. This is a little less than the world population, but it's a nice round number.

3.1 Random Networks

Suppose first that everyone has random friends. In the way we commonly think of structure, we might say that this network has none. We might also say that the network is not complex in that it can be defined simply: give each person 150 random friends. However, these insights are not quite correct. The reason is that the fact that friendships go in both directions imposes structure, though not much and increasingly little as the number of people gets large.

We also might think of this network as maximally diverse, since any two people should expect to have almost no overlap in their neighborhoods or in their two-neighborhoods when the number of people is huge. We begin by making a crude calculation of the growth in the size of the k -neighborhoods.

Random Network: Neighborhood Sizes (crude calculation)

Neighborhood Size = 150

Two-Neighborhood Size = 22,500

Three-Neighborhood Size = 3,375,000

Four-Neighborhood Size = 506,250,000

Five-Neighborhood Size = 75,937,500,000

According to this calculation, most people would have seventy five billion people in their five neighborhood. Given that there are not that many people in the world, you should realize that our approximate calculation at some point became very crude. The problem is that we're ignoring the possibility of one of my three neighbors also being one of my two neighbors or of the same person being a three neighbor by two different means. It's not too difficult to correct for this overlap but we won't bother doing so here.

3.2 Cliques

In the next network structure we consider, people belong to cliques of size one hundred and fifty one. Every pair of people within a clique are connected and no person is friends with anyone outside her clique.

Cliques Network: Neighborhood Sizes

Neighborhood Size = 150

Two-Neighborhood Size = 150

k-Neighborhood Size = 150

Here we see the other extreme. Neighborhood sizes stay constant. The world consists of a congeries of isolated cliques. The number of degrees of separation between two people in distinct cliques is infinite. Again, here, we might think that this is an unrealistic assumption. However, there are some that would argue that the Internet could create greater balkanization of social networks (see Van Alstyne and Brynjolfsson (1996)). Straightforward intuition underpins this concern. If I can interact with anyone that I want and if I prefer to hang out with people with similar interests, then we may all sort into little clusters of like minded people. This may not happen in the social realm given our diverse interests but it could happen in the scientific arena where those people interested in Algebraic Topology can now free themselves of their physical locations and relocate into virtual isolation. If this happens, the scientific world could evolve into a congeries of isolated cliques.

3.3 Small Worlds Networks

The final class of networks we consider are called small world networks. A small world network is a combination between a random network and a clique network.¹ People have a substantial interconnected group of friends and then they have some random friends as well. There is ample empirical evidence that social networks have small world type features. This combined with the mathematical tractability makes them an ideal technical tool.

We will model this as follows: Each person has 140 friends who belong to a clique and ten random friends. It is not hard to approximate the growth in the neighborhood sizes given these assumptions.²

Calculating the size of the small world k -neighborhoods is made simpler if we adopt the following convention: let C denote the friends in a clique and R denote the random friends. Let CR denote the random friends of the friends in the clique, and CRC denote the friends in the clique of the random friends of the initial clique friends, and so on. Using this notation $CC = C$. The clique friends of people who belong to the clique also belong to the clique. Therefore, the k -neighborhood consists of all strings of R 's and C 's of length k which do not have two consecutive C 's.

Small Worlds Network: Neighborhood Sizes (crude calculation)

Neighborhood Size = 150:

¹In a recent book, Duncan Watts provides a detailed mathematical analysis of small worlds models. Among other things, he proves theorems about minimal paths and degrees of separation.

²These calculations are crude in the sense that in calculating the size of the $k + 1$ -neighborhood, we should often be multiplying by 149 instead of 150 to account for the friend from the k -neighborhood who is already counted as well. Further, in many cases where we assume ten random friends we should really assume nine.

$C = 140$ and $R = 10$

Two-Neighborhood Size = 2900:

1400 CR, RC
100 RR

Three-Neighborhood Size = 239,000:

196,000 CRC
14,000 $CRRRCRRRC$
1000 RRR

Four-Neighborhood Size = 6,450,000:

1,960,000 $CRRC, CRRCR, RCRC$
140,000 $CRRR, RCRR, RRCR, RRRC$
10,000 $RRRR$

Five-Neighborhood Size = 399,100,000:

274,400,000 $CRCRC$
19,600,000 $CRRCR, CRRRC, CRCRR, RCRCR, RCRR, RRCRC$
1,400,000 $RRRRC, RRRCR, RRCRR, RCRRR, CRRRR$
100,000 $RRRRR$

Six-Neighborhood Size = 4,789,000,000:

2,744,000,000 $CRCRCR$
196,000,000 $CRCRRR, CRRCRR, CRRRCR, CRRRRC, RCRCRR, RCRRCR, RCRRRC$
14,000,000 $CRRRRR, RCRRRR, RRCRRR, RRRCRR, RRRRCR, RRRRRC$
1,000,000 $RRRRRR$

The increase in the neighborhood sizes from the small worlds network are not as large as in the random network. We would not expect them to be. What is surprising though is that only a limited amount of network diversity is needed to generate connectedness quickly. If each person had only ten friends, it would take nine degrees of separation to get to one billion people. What happens in the small worlds model is that the friends of the friends from the clique also expand into a large network of people, so a little diversity goes a long way.

In looking at these numbers we cannot help but realize why Grannovetter got the result that he did. If I have 150 friends and 3000 friends of friends but a quarter of a million friends of friends of friends, who is more likely to help me find a partner or a job or to tell you a funny story? Based on their number, the friends of friends of friends would seem to be the most likely group. We can formalize that intuition with an example. Suppose that in a given month that the odds of running into a given friend and having that friend introduce you to a potential date is one in a one thousand. Assuming independence across events the odds of a friend introducing you to someone is about 14%. Suppose that the odds of running into a given friend of a

friend of a friend and that friend of a friend introducing you to someone are a mere one in a million. In this case the odds are 22% that you'll meet a potential date even though friends are one thousands times more likely to introduce you to someone than are friends of friends of friends. Nevertheless, the sheer number of friends of friends of friends outweighs their low probability of helping.

3.4 A General Model

We can now create a single model in which the three models we have already discussed can all be considered special cases. Consider a generalization of our small worlds model in which C and R can vary. A moments reflection will convince you that if we let $C = 0$, then we have the random model. Each person has 150 random friends. And, if we let $C = 150$, then each person has only friends in their clique and we have the clique model.

If the 3-neighborhood is important for getting jobs, finding partners, creating social capital, and sharing political information and gossip, then perhaps we should ask how the three neighborhood grows as a function of C and R . From above we know that the total number of people in the three neighborhood equals $CRC + CRR + RCR + RRC + RRR$, but this can be reduced to $CRC + 3RRC + RRR$ which can be further simplified with a few tricks to $150 * R * (150 + R) - RRR$. We can calculate this for a few values of R

R	Size of 3- Neighborhood
1	22,649
2	45,592
3	68823
4	92336
5	116125
6	140184
7	164507
8	189088
9	213921
10	239000
11	264319
12	289872
13	315653
14	341656
15	367875

If we look at these numbers we see that the increase in the size of the 3-neighborhood is not exponential. It is a geometric function with a large linear term. If we look at the formally carefully for small R , we see that it is approximately $150 * 150 * R$. In the table below, we see the accuracy of this approximation.

R	Size of 3- Neighborhood	$150 * 150 * R$
1	22649	22500
2	45592	45000
3	68823	67500
4	92336	90000
5	116125	112500
6	140184	135000
7	164507	157500
8	189088	180000
9	213921	202500
10	239000	225000
11	264319	247500
12	289872	270000
13	315653	292500
14	341656	315000
15	367875	337500

The key insight from this calculation is that if we have maximally diverse neighborhoods then initially the rate of growth of in the size of our k -neighborhoods is exponential. *Greater diversity in connections implies faster neighborhood size growth.* This example is illustrative because people tend to toss out the word “exponential” every time something grows fast. But this function is only geometric. Moreover, for the R that interest us most, the linear term is dominant.

3.5 Locating Minimal Paths

Our general model of networks as formulated has several flaws, but one is particularly glaring. In real networks, people can find paths between themselves and others quickly. We do this by matching features of the target person with features of our friends. If I wanted to make a personal connection to a doctor in Fairbanks, I would search paths in my personal network beginning with Alaskan friends or with friends who were doctors. In the stark model described so far this would be difficult for someone. The person would have to exhaustively search successive k -neighborhoods until connecting with the desired target person. In recent work, Strogatz (2001) has shown that if you place people geographically and then make the distribution of random friends geographically biased near the person, the minimal paths can be located relatively easily. The idea is that someone from California’s random friends would be biased toward the west coast. Therefore, if you wanted to connect to someone in Oregon and you had a random friend in California and the rest were east of the Mississippi, you’d search for a path beginning with the Californian.

4 Multiple Paths and Robustness

In 1994, Southern California was rocked by the Northridge Earthquake. The highway system suffered massive damages. In preparing for such an eventuality the state and city had prioritized bridges according to how badly they needed of repairs. The problem with this approach is that when the earthquake hit, the resulting damage looked like strategic bombing. Almost every major North-South and East-West route was compromised. Alternatively, they might have prioritized freeways. If they had, instead of having say an earthquake which destroyed six overpasses wipe out five freeways, they might have had an earthquakes which destroyed eight overpasses but only wiped out two freeways.

Even after the damage caused by the earthquake, the massive traffic flows in Los Angeles continued, though not quite unabated. They were able to do so because there were multiple paths between locations – there were lots of routes between Pasadena and Santa Monica. Traffic was just diverted. In contrast, when the pass between Yosemite and Tahoe closes for the winter, traffic just stops. The only alternative routes involve hundreds of miles of travel.

We can apply this same idea to ecology, economics, organizational theory, and information theory.

4.1 Ecology: Food Webs

In a food web, the objects are species and a connection implies that one species eats the other. If we follow a food web down to its final nodes, we find species that get their energy directly from the soil, freshwater, sea water or the sun, or some combination of these things. Species further up the food chain consume the same energy but it proceeds up a path. Almost all species have multiple paths to that energy. If not, it would mean that each species in the food web only ate one other species. This would make for an extremely fragile system.

4.2 Economics: Supply Chains

A house or a car consists of many component parts. These parts are provided through supply chains. If a firm has a unique path to get a part they suffer two consequences. First, breakdown. If something along the path collapses, the firm can be forced to shut down. Second, they may have to pay high prices for that part as the suppliers along the chain will realize that they have market power.

4.3 Power Laws and The Internet

Some networks have a power law distribution in connections. Under a power law distribution, if we arrange the nodes in order according to the number of links, then we get an exponential function. This means that there will be one node with an

enormous number of connections and a handful of other nodes with large numbers of connections.

This distribution will arise under many circumstances. For example, if new links connect to a node in proportion to the number of links a node already has then a power law will result (See Barabasi's book *Linked*). A large portion of the Internet can be reasonably approximated by a power law distribution.

The distribution of connections from the nodes plays a role in the robustness of the network. Suppose that in order to be robust, there must be a path from any node to any other. In this formulation, a robust network is equivalent to a connected network. Suppose that we have random failures of nodes. With a power law distribution of connections, these failures will not be likely to make the graph disconnected. Why? Because most nodes are only connected to a couple of other nodes. The only way that random failures would cause the graph to lose connectedness would be for those random failures to hit the handful of most connected nodes. A highly unlikely event.

However, if the failures are not random but directed, then the most connected nodes can be wiped which is likely to compromise the system. A directed failure can be thought of as an attack. Therefore, on the one hand, if we have diversity in the connectedness of nodes then the network is more robust to random failures but less robust to attack. Here then is even more evidence of why we have to be careful when saying "diversity implies robustness."

4.4 Organization and Information Theory: Robustness

The applications of the multiple paths idea to organization and information theory are similar. In each case, multiple paths prevent error in message passing.³ To see how this works, consider a variant of the telephone game. Suppose that someone at level one hears a number. They repeat that number to someone at the next level. Suppose that after proceeding up the chain there is ten percent chance that the number is decreased by one and a ten percent chance it is increased by one. If the true number was ten, then we can write the probability distribution over the number heard as follows:

number heard	9	10	11
Probability	10%	80%	10%

Now suppose that there are three independent paths that can get the information up through the organization. The totals for those three messages will be between 27 and 33. The probabilities for each of these numbers are in the table below, as well as the best guess given each sum. For example, if the numbers add to 28, 9 is the best guess, but if they add to 29, 10 is the best guess.

³See Jonathon Bendor's book *Parallel Systems*.

Sum of numbers heard	27	28	29	30	31	32	33
Best Guess	9	9	10	10	10	11	11
Probability	0.1%	2.4%	19.5%	56.0%	19.5%	2.4%	0.1%

With three independent paths, the information will be correct 95% of the time. Thus, *multiple paths imply greater robustness*, at least in the context of information flow over networks.

5 Strategic Choice of Connections

We now analyze the strategic choice of connections and ask what sort of networks are likely to emerge. This is a fascinating and complex topic. Agents, be they people, firms, species, organizations, or anything else have preferences or desires over not only what they are connected to but also over the entire structure of the network. Unfortunately though, in most cases agents only have control over who they connect to in the network and not over higher order neighborhoods. And, in many cases, such as predator prey networks, agents may not have much say even over who or what connects to them. We will consider three approaches to understanding the incentives for creating connections.

5.1 Weak and Strong Ties

In our general model with clique and random friendships, we showed that the number of friends of friends of friends grew rapidly with the number of random friends. This means that there is an incentive for agents to connect to random friends rather than clique friends if they want to get more information. The increase in friends of friends of friends is not as dramatic as might be suggested by the earlier model though in that that model assumes that everyone increases their number of random friends not just an individual agent. Nevertheless, it is in the interest of agents to make more weak ties, what we call random friends, if they want to maximize the amount of information they obtain (see Matt Jackson's web page for a more complete analysis).

Why then wouldn't agents have all random friends then? After all, this would be the best way to get lots of information. If we just go back a few lectures, we will see that we have already answered this question. We don't make friends for the sole purpose of getting lots of information. We have multiple and layered reasons for connecting with others. People prefer to hang out with people with similar interests. Therefore, we should expect to see clusters of like minded friends.

5.2 Tassier and Menczer

Making random connections to get more information may be in a person's best interest but it may not be in the interests of that person's group. If one group has greater

control over resources and information then the group may collectively suffer if its members make random connections to members of the other group. This idea is due to Troy Tassier and Filippo Menczer who show that more segregated networks enable agents to keep information within a community. Their result overturns the standard small world logic that we want to be as connected to everyone as possible. In fact, if one group has greater control over information or resources, that group's members want to do the opposite and restrict the number of connections to outside sources.

5.3 Structural Holes

In making personal connections, we choose people we like, but we also choose people who may have information, knowledge or ideas that will be useful to us. In forming work relations, this latter influence may be predominant. The best connections to make often depend on the connections that other people make. To capture this idea, Ronald Burt came up with the notion of a *structural hole*. Burt argues that successful people fill structural holes. A structural hole is not unlike a good niche – it's a location that reaps big benefits.

When someone fills a structural hole, they make a connection that makes the network more efficient. Suppose for example that the an administrator wants to know what students think. Suppose he also wants to get in shape. He could go to the gym with one of his deans and ask that the dean find out from her professors who in turn ask the students how they feel. This has some obvious problems, as our previous mathematical example suggests. Alternatively, some student on a lark could send the president an email saying “would you like to join our 10pm basketball league?” The administrator might realize that this fills two of his needs and accept. The students, in this case, will have filled a structural hole, and in doing so have access to the administrator and can influence on how the university is run.