Why your friends have more friends than you do. The exciting world of random networks.

Winfried Just
Department of Mathematics
Ohio University

Athens, Ohio, October 24, 2012

The number of friends of your friends

Let v be a person, and let $d(v)$ denote the number of v's friends.

Let $F(v)$ denote the set v's friends, and let $d_1(v)$ be the arithmetic mean of the set $\{d(w) : w \in F(v)\}$.

I claim that **on average** $d(v) < d_1(v)$.

This is an outrageous claim, for at least two reasons:

- It seems counterintuitive. Since we have made no special assumptions about v or v's friends, it seems that **on average** v should have about as many friends as v's friends have **on average.**
- I (or anybody else) has only very little knowledge about the actual number of friends of other persons.

At least I'm in good company

The title of my talk is actually taken from a famous journal paper that appeared two decades ago:

Feld, Scott L. (1991), "Why your friends have more friends than you do", *American Journal of Sociology* 96(6): 1464–1477.

In the paper, the author gives a mathematical proof of my outrageous claim.

How can one mathematically prove any such thing?

First we need to model people's friendships with suitable mathematical structures.

Graphs

A **graph** consists of a set V of **vertices** or **nodes** and a set E of**edges** that connect some of the nodes. Formally, an edge $e \in E$ is an unordered pair $\{v, w\}$ of distinct nodes that the edge e connects.

See `http://en.wikipedia.org/wiki/Gallery_of_named_graphs` for some nice pictures of graphs.

For example, the friendships between a group of people V can be modeled by a graph whose nodes are the people in this group, and an edge $\{v, w\}$ signifies that v and w are friends.

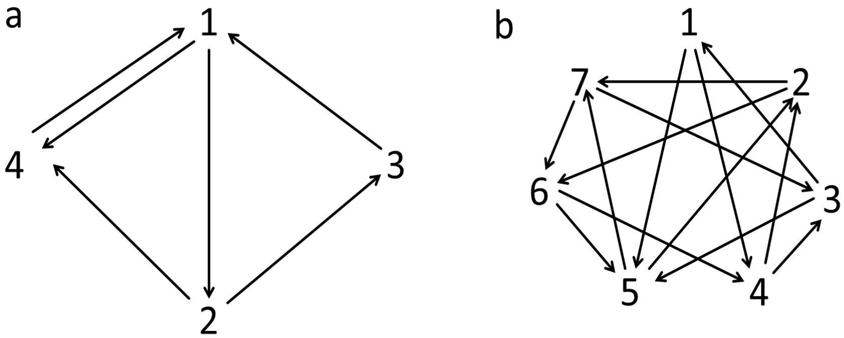
The **degree** $d(v)$ of a node v is the number of edges that connect to v. Note that in the friendship graph this is exactly the number of v's friends.

Digraphs

Friendships are (usually) symmetric. In many situations the connections between nodes have a direction, and for those the model of choice is a **directed graph** or **digraph,** where instead of edges we have a set A of **arcs.** Formally, an edge $a \in A$ is an ordered pair (v, w) of distinct nodes that the arc a connects.

In a directed graph, each node has two types of degrees, the **indegree** $ind(w)$ of a node w is the number of arcs that target v and the **outdegree** $out(v)$ of a node v is the number of arcs that originate from v.

Two examples of digraphs



For example, in digraph (a) node 4 has outdegree 1 and indegree 2.

Networks

Let us call a mathematical structure a **network** if it is either a graph or a digraph. The edges or arcs will be collectively referred to as **links.** Here is a small sample of real-world structures that can be modeled as networks:

- Friendships between people (graph)
- Sexual contacts between people (graph)
- The World Wide Web (digraph, with arc (v, w) signifying that there is a hotlink to page w at page v)
- Neuronal networks, aka brains (digraph, with arc (v, w) signifying a synaptic connection)
- Transportation networks (usually graphs)

Some common features of these networks

- The number of nodes is very large (billions of people and web pages, about a trillion neurons in the human brain).
- The network keeps changing over time.
- The connectivity (the set of links) at any one time cannot be fully known.
- However, we can get some idea about the network by probing the connectivity of a subset of its nodes.
- In particular, we can get usually get reasonably good estimates about the **degree distribution** in these networks, that is, of the proportion P_k of nodes that have degree k, where k is a nonnegative integer.

Some challenges in collecting data on networks

Suppose you want to empirically study the degree distribution in the friendship network and the network of sexual contacts. Consider the following questions. Which ones have clear-cut answers? For which ones should you expect an honest answer?

- Would you feel comfortable telling us how many friends you have?
- How many friends do you have?
- Who, exactly, counts as a friend?
- So, how many friends do you have?
- Who, exactly, counts as a sex partner?
- Do you remember the total number of your sex partners?
- So, how many persons did you ever have sex with?

Some common features of these networks

- The number of nodes is very large (billions of people and web pages, about a trillion neurons in the human brain).
- The network keeps changing over time.
- The connectivity (the set of links) at any one time cannot be fully known.
- We **may** get some idea about properties of the network, in particular about its degree distribution, by probing the connectivity of a subset of its nodes.

In view of these uncertainties it makes sense to study **random networks** that satisfy certain **structural assumptions,** in particular, assumptions about the degree distribution. Since there usually is some randomness in the process of making connections between the nodes of networks of interest, one may hope that the mathematical conclusions carry over to the real networks.

The expected number of your friends

Assume that the friendship network is a **random graph** with a fixed set V of nodes. Then both $d(v)$ (the number of v's friends) and $d_1(v)$ (the mean of $d(w)$ for v's friends) are random variables.

For each nonnegative integer k let p_k denote the probability that a randomly chosen v has exactly k friends.

Then the mean value of $d(v)$ is equal to

$$\mu = E(d(v)) = \sum_k kp_k.$$

The expected number of friends of your friends

To calculate the expected value of $d_1(v)$, let us pick v randomly and let w be a randomly chosen friend of v.
As long as friendships are being formed sufficiently randomly, the probability of w being chosen in this way is equal to $\frac{kp_k}{\mu}$, where k is the degree of w. Thus

$$E(d_1(v)) = \frac{1}{\mu}\sum_k k^2 p_k == \frac{E((d(v)^2)}{\mu} = \frac{Var(d(v)) + \mu^2}{\mu}.$$

The right-hand side of the above is larger than μ as long as $Var(d(v)) > 0$.

That is, as long as there is any variability in the degrees of individual nodes, we will have

$$E(d_1(v)) > E(d(v)).$$

But wait a minute ...

How can the above argument possibly be true?
What about a situation when each person has either exactly one friend (is a loner) or exactly 100 friends (is gregarious) and loners befriend only other loners? Then clearly $d(v) = d_1(v)$ for all v and the above result fails, although there is variability in the degrees.

A network like this is called **completely assortative.**

Recall that we assumed that **friendships are being formed sufficiently randomly,** which **in this case** means that there is no dependence between $d(v)$and $d(w)$ if we know the $\{v, w\}$ forms an edge.

This brings us to a more general question:

What is a random network anyway?

Erdős-Rényi networks

Of course mathematically treating a network as "random" requires specifying a probability distribution, at least implicitly, on the class of all networks. The easiest approach is the following:

- Fix the set V of nodes of size $|V| = n$.
- Fix a probability p.
- Include each potential link (edge $\{v, w\}$ or arc (v, w)) in the network with probability p, independently for different links.

This construction implicitly defines the class of **Erdős-Rényi networks.**

It is easy to see that this construction gives networks with mean degree (indegree, outdegree) $p(n-1) \approx pn$. The standard deviation of the degrees in these networks is of order $\sqrt{n}$, which implies that for the vast majority of nodes v, w we should expect that $\frac{d(v)}{d(w)} \approx 1$. This, however, is not what we observe in most real-world networks of interest.

Scale-free networks

Empirical studies show that many large networks of interest have degree distributions that roughly follow a **power law** with $p_k = \alpha k^{-\gamma}$ for some $\alpha, \gamma > 0$.

Such networks are called **scale-free networks.** Most nodes form very few links, but a few nodes, the so-called **hubs,** have a huge number of links.
Think of all the hotlinks pointing to YouTube *vs.* the ones pointing to my own homepage.

As we have seen, Erdős-Rényi networks are not expected to be scale-free. How can one randomly generate networks whose degree distributions obey a power law?

The preferential attachment model

One possibility is the following.

- Grow the network by adding one node at a time, by adding a new node v_n to the current network N_n.
- Let the probability that v_n gets linked to $w \in N_n$ be equal to $f(n, d(w))$, where f increases in its second argument $d(w)$.

It can be shown that this procedure will,for suitably chosen f, give networks with an expected power law distribution of the degrees.

Note that in this scheme already highly connected nodes attract more new links. This has been called "the rich get richer phenomenon."

It seems to be the way in which the www and friendship networks are evolving.

Dynamical systems on networks

Nodes in a network often can exhibit distinct **states** that change over **time.** The change of the state of the entire network over time is called **network dynamics.**

- Nodes in transportation networks can be congested at times.
- Neurons can fire or be at rest.
- People can be sick or healthy.

Note that in all of the above examples changes in the state of a node occur in response to interactions along the links of an underlying network.

How does the network connectivity influence the network dynamics?

Example: A model of certain brain structures

The following is true in at least some neuronal networks.

- Neurons fire or are at rest.
- After a neuron has fired, it has to go through a certain **refractory period** when it cannot fire.
- A neuron will fire when it has reached the end of its refractory period and when it receives firing input from a specified minimal number of other neurons.

One can build a simple model of neuronal networks based on these facts.

A model of neuronal network dynamics

Ahn S, Smith BH, Borisyuk A, Terman D, Physica D, 2010

A directed graph D and integers n (size of the network), p_i (refractory period), th_i (firing threshold).

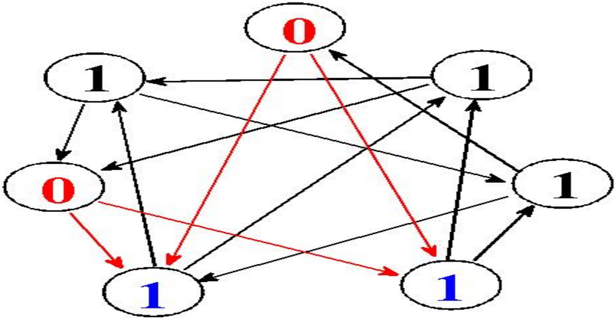
A state $\vec{s}(t)$ at the discrete time t is a vector:
$\vec{s}(t) = (s_1(t), \ldots, s_n(t))$ where $s_i(t) \in \{0, 1, \ldots, p_i\}$ for each i.
The state $s_i(t) = 0$ means neuron i fires at time t.

Dynamics on the discrete network $N =< D, \vec{p}, \vec{th} >$:

- If $s_i(t) < p_i$, then $s_i(t+1) = s_i(t) + 1$.
- If $s_i(t) = p_i$, and there exists at least th_i neurons j with $s_j(k) = 0$ and $< j, i > \ \in A_D$, then $s_i(t+1) = 0$.
- If $s_i(t) = p_i$ and there do not exist th_i neurons j with $s_j(t) = 0$ and $< j, i > \ \in A_D$, then $s_i(t+1) = p_i$.

An Example

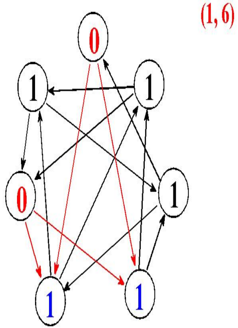
Assume that refractory period= 1 and threshold= 1.

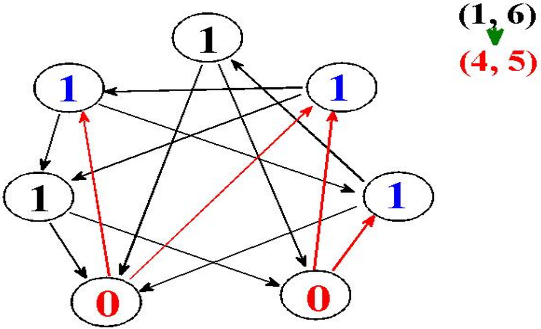


(1, 6)

An Example

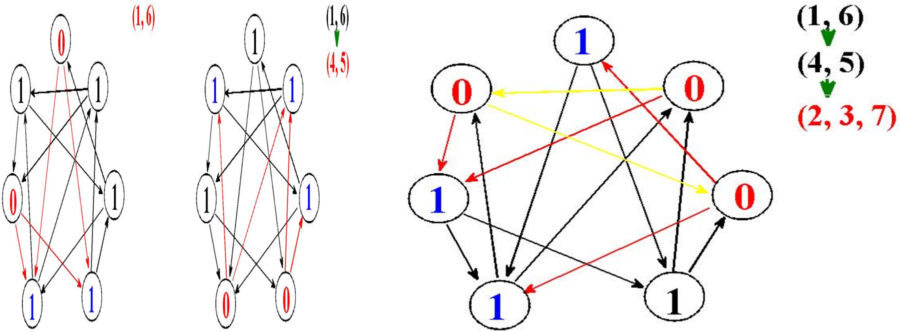
Assume that refractory period= 1 and threshold= 1.





An Example

Assume that refractory period= 1 and threshold= 1.



Two phase transitions by Just W, Ahn S, Terman D, Physica D. 2008

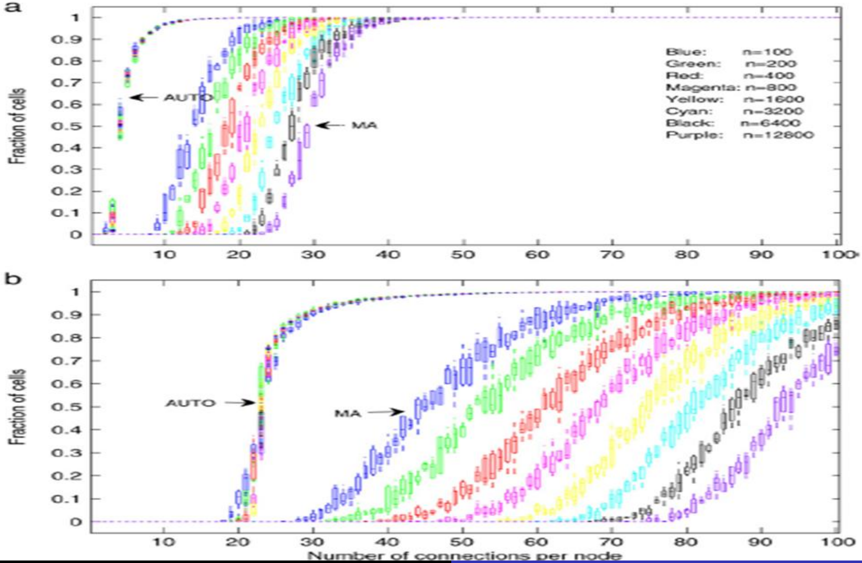
What happens if D is an Erdős-Rényi random digraph with n nodes such that each potential arc is included with probability $\rho(n)$?

Theorem

1. *A first phase transition occurs at $\rho(n) \sim \frac{\ln n}{n}$: Above this threshold, in a generic trajectory all nodes will always fire as soon as they reach the end of their refractory period.*
2. *A second phase transition occurs at $\rho(n) \sim \frac{c}{n}$: Above this threshold, in a generic trajectory most nodes will always fire as soon as they reach the end of their refractory period.*

Results of the simulations Just W, Ahn S, Terman D, Physica D. 2008

How does a "phase transition" look like?



Some directions for further explorations

- Another phase transition was detected for $\rho(n) \sim \frac{1}{n}$. Together with S. Ahn and A. Borisyuk we are currently exploring what goes on in this (most interesting) range of connectivities.
- What happens for random digraphs other than Erdős-Rényi random digraphs (e.g., the ones obtained from the preferential attachment model)?
- The book chapter "Neuronal Networks: A Discrete Model," Just W, Ahn S, Terman D; to appear in *Mathematical Concepts and Methods in Modern Biology,* Robeva R and Hodge T eds., Elsevier, January 2013

  includes eight research projects that gradually lead students from relatively easy exercises to unsolved open problems. Most of these are at a level that is accessible to undergraduates.

Disease dynamics

Let us consider a different problem of applied mathematics: Modeling the spread of infectious diseases.

One basic model is the so-called **SIR model.** In this model, the population is partitioned into three classes:

- S comprises all individuals that are **susceptible** to infection.
- I comprises all individuals that are **infectious**.
- R comprises all individuals that are **removed**.

In this model one assumes that removal occurs by recovery from the disease or by death and confers permanent immunity to reinfection.

Differential equations for the SIR model

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI,\\ \frac{dI}{dt} &= \beta SI - \gamma I,\\ \frac{dS}{dt} &= \gamma I.\end{aligned}$$

The model predicts that if a small number of infected individuals is introduced into a susceptible population, then

1. either the disease will quickly die out with a negligible proportion of individuals becoming infected,
2. or it will become an epidemic and die out only after a fixed fraction of the population has become infected and been removed.

Which property makes the difference between these two scenarios?

The parameter R_0

The **basic reproductive ratio** R_0 is defined as

“the **average** number of secondary cases arising from an **average** primary case in an entirely susceptible population.”

Theorem

If $R_0 < 1$, the disease will quickly die out, if $R_0 > 1$, an epidemic will result.

Proof: Suppose that k infected individuals are introduced into and otherwise entirely susceptible population of n individuals. If $k \ll n$ and T is the average time an individual stays infectious, then for small enough m, after time mT we will have

$$E(|I|) \approx kR_0^m.$$

Contact networks and disease dynamics

Notice that in reality, an infectious disease is usually spread during a **contact** between an infectious and a susceptible individual.

Thus the actual spread of a disease will be influenced by the structure of the relevant contact network (*e.g.* sexual partners, friends, co-workers, commuters), with a given contact resulting in a transmission with a certain probability.

The proof I just gave you makes a lot of hidden assumptions; in particular, it assumes that the contact network is not assortative by degree. This is not a realistic assumption for most actual contact networks.

In particular, for the same value of R_0, which was defined in terms of averages, in a scale-free network that is somewhat assortative by degree, an epidemic might result if the initially infected individuals are hubs, while the disease may quickly die out if the initially infected individuals have relatively few contacts.

An exciting area of research

This leads to a more general questions:

Which properties of the underlying contact network determine the spread of an infectious disease?

Given our knowledge about the contact network, how should interventions be targeted so that they will most effectively prevent epidemics?

In collaboration with Dr. Grijalva from the College of Osteopatic Medicine and my Ph.D. student Bismark Oduro we are investigating this type of questions for Chagas' disease, which is endemic in South and Central America.

A current epidemic

Over the last decade or so, the average researcher in the area of large networks and their dynamics has infected an average of $R_0 > 1$ susceptible colleagues or students with this type of interest. The number of researchers working on this topic has grown exponentially and we have an epidemic. Should you worry?

Symptoms include:

- A state of heightened excitement about one's research that can lead to sleep deprivation.
- Increased susceptibility to authorship in prestigious journals.
- Increased risk of exposure to funding agencies such as NIH.
- Increased risk of contamination by genuine real-world applications.
- Increased exposure to the lure of nonacademic employers.
- So far, not many cases of full recovery have been observed.