Interpret **network** mathematically as (directed or undirected) **graph**. Recall a graph has **vertices** (= nodes, sites) and **edges** (= links, bonds). Interested in case

- $n =$ number of vertices is large

- $\alpha =$ average out-degree (degree = # edges at a vertex) is much smaller than n

- network has "just grown" without centralized planning.

The Internet is obvious contemporary example; also transportation networks (railroads, highways, air routes); telegraph, telephone, electrical, water supply. Many such networks are intrinsically 2-dimensional; Internet interesting because it's not ("The Death of Distance").

To a statistician, a large network is merely a particular kind of large data-set. Two philosophically different ways to think about data:

- descriptive statistics; seek to summarize in human-interpretable way
- probability models; seek some procedure (rules + randomness) to synthesize data similar to real data.

Mathematically, these are not so different.

Freshman example. Given data on heights and weights of 1000 people, we can summarize these 2000 numbers by 5 numbers

ave and s.d. of height
ave and s.d. of weight
correlation coefficient.

This choice of summary statistics corresponds perfectly to a certain (bivariate Normal) probability model. Saying the model fits the data is the same as saying there is no "structure" in the data beyond what the summary statistics describe.

Unfortunately . . . . . . no-one knows how to choose descriptive statistics or probability models well for complex networks.

I'll talk about 2 areas of great current activity.

2. "The WWW graph" of web sites and links, studied by e.g. the Self-Organized Networks group at Notre Dame Physics dept. Claimed empirically

$$P(\text{out} - \text{degree} = i) \propto i^{-\gamma_{out}}; P(\text{in} - \text{degree} = i) \propto i^{-\gamma_{in}}$$

for some **scaling exponents** $\gamma_{out} \approx 2.45, \gamma_{in} \approx 2.1$.
Seek a model to explain this.

3. Small world social networks: vertex = person, edge = friends (or other relationship). Claimed empirically that distance Δ between two typical people is small. Seek a model to explain this.

1. What mathematicians know about "locally tree-like" random graphs.

The Erdős - Rényi model

Mental picture: n vertices, initially no edges, add edges between random pairs of vertices until prescribed average out-degree α reached.

Basic properties, for typical vertex v, in the relevant case $\alpha > 1$.

(1) out-degree and in-degree are independent, Poisson(α) distribution.

(2) $B(v, d) =$ number of vertices within d edges of v grows $\approx \alpha^d$.

(3) Locally tree-like: no small cycles thru v.

(4) Distance between two typical vertices $\approx \log n$.

This basic model has been very intensively studied by theorists; never intended to model real networks. First unrealistic feature: degree distribution of real networks not Poisson. But one can modify the model (in several different ways) to obtain more general degree distributions.

rod-ball model. Prescribe joint distribution $q(i,j) = P$(in-degree $= i$, out-degree $= j$).

$$\bigcirc \quad \bigcirc \quad \bigcirc \quad \bigcirc \quad \bigcirc \quad \bigcirc \quad \bigcirc \quad \bigcirc$$

Sample without replacement edge-pairs and make one edge.

Weighted-vertex model. Prescribe distribution of "weights" $(w_{in}(v), w_{out}(v))$, normalized to have averages $\bar{w}_{in} = \bar{w}_{out} = 1$. Then make

$$P(\text{edge } v \to v') = \alpha w_{out}(v) w_{in}(v')/n.$$

This allows us to specify the joint distribution of in/out degrees fairly generally (subject to being more-spread-out than Poisson).

Folklore: these models share the other properties (2 - 4) of the basic random graph model.

Why a scaling law for WWW degree dist.?

Here's a simple possible explanation. The assertion
(1) Number of WWW sites is growing exponentially at rate ρ

is equivalent to:

(1*) A random site has existed for random time T, where $P(T > t) = e^{-\rho t}$.

Another plausible notion:

(2) Number $N(t)$ of links into a site at time t after its creation has $N(t) \approx e^{\beta t}$, some $\beta < \rho$.

Now (1) and (2) immediately imply that the number X of links into a random site has

$$P(X > i) \approx i^{-\rho/\beta}.$$

This doesn't mention an explicit random graph model, but indicates that any model consistent with (1) and (2) will have a scaling law. In other words, a scaling law is a natural consequence of exponential growth – older sites have more links – without explicit modeling of "importance".

Snakes: observed and calculated numbers of genera of each size. (Table VII of Yule (1924)).

Number of species in genus	Number of genera: Observed	Number of genera: Calculated
1	131	130.9
2	35	47.2
3	28	25.2
4	17	16.0
5	16	11.2
6	9	8.3
7	8	6.5
8	8	5.2
9 to 11	13	11.1
12 to 14	3	7.2
15 to 20	7	8.8
21 to 34	14	9.2
35 upwards	4	6.2
Total	293	293.0

Yule (1924) proposed a model of evolution to explain the scaling law here.

- Start with 1 genus with 1 species.
- From each species, at constant (stochastic) rate β, a new daughter species in the same genus arises.
- From within each genus, at constant (stochastic) rate ρ, a new species of a different genus arises.

This completely specifies a model with exponential growth of number of species and of number of genera, and leads to an explicit 1-parameter family of distributions for

$$X = \text{number of species in a genus}$$

with $P(X > i) \approx i^{-\rho/\beta}$.
Again, scaling arises from exponential growth; older genera contain more species.

[Biologists never accepted this explanation.]

Barabási - Albert (*Science*, 1999) proposed the following model for WWW links.

- each site has a fixed number m of out-links, and a varying number of in-links.

- when site first created, the m out-links go to m sites chosen at random but with probabilities proportional to their total number of links.

Easy back-of-an-envelope calculation. Write $t =$ number of sites. Consider $F(x, t) =$ number of sites at time t with $> x$ links; then write down and solve

$$\frac{d}{dt}F(x,t) \approx -\frac{x}{2t}\frac{d}{dx}F(x,t)$$

to get long-time distribution of number of links at a typical site

$$P(X = i) \approx 2m^2 i^{-3}, \; i \geq m.$$

Bianconi-Barabási (2000 preprint) discuss a "weighted" variation:

- each site v has a weight $w(v)$ chosen from a weight-distribution W

- when a new site is first created, the m out-links go to m sites v chosen at random but with probabilities proportional to $w(v)\times$ (total number of links at v).

This gives a scaling law

$$P(X = i) \approx i^{-\gamma}$$

for γ depending on the distribution W;
$\gamma = 2.255$ if W has $U(0, 1)$ distribution.

All these models are locally tree-like. To model "social networks", it's relevant that two people who are both your friends are quite likely to be friends of each other. Can define **clustering coefficient** of a graph:

$C =$ proportion of 2-paths which extend to a 3-cycle

Lattice and tree-like models fail to reproduce C for opposite reasons. Recent literature on **small world models** surveyed by Newman (2000) seeks to invent and study models with

- local clustering
- and small diameter.

Recall vertex = person, edge = some type of social relationship. One version of **small world model** (Newman - Watts 2000) is:

- Take the usual d-dimensional cube Z_L^d but include edges to all k nearest neighbors

- each site, with some small probability p, is also linked to a (uniform) random site via a **shortcut**.

Study in-graph distance Δ between two typical vertices; argue that in $d = 1$

$$E\Delta \approx \frac{\log(2Lp)}{4kp}, \quad pL \gg 1.$$

Also studied non-rigorously: spread of information/epidemics; connectivity. Barbour - Reinert (preprint 2001) give careful analysis of a continuum version.