Networks

14th October 2013

Issues and questions

- Historically, technical progress appears to be strongly linked to industrialization.
- The latter has emerged in specific places and at specific times.
- This fact makes technical progress idiosyncratic: it is context-specific, appearing in given industrial sectors; firm-specific.
- Where it has emerged it has shown to be cumulative and unfolding along specific trajectories.

Some characteristics

- Technical progress is quite obviously knowledge-based.
- More to the point, it appears to be driven by learning and searching in various forms.
- For instance: learning-by-doing and learning-by using.
- All these activities require information: knowledge and learning build upon information
- The context in which this happens seems to indicate that clustering of firms, of industrial sectors, of population settlements has, in fact, occurred.

Some related phenomena

- The rise of markets and related urban centres.
- The growth of cities: some have grown to be very large, others have remained small. Clearly, within a context of 'rise and fall'.
- Flows of information are crucial for learning and searching to come to pass as well as for trade to flourish.

An important stylized fact

- The distribution of city sizes.
- The distribution of scientific paper and of patent citations
- The distribution of papers written by scientists
- The distributions of WWW web pages

$$f(x) \sim x^{-\gamma}$$

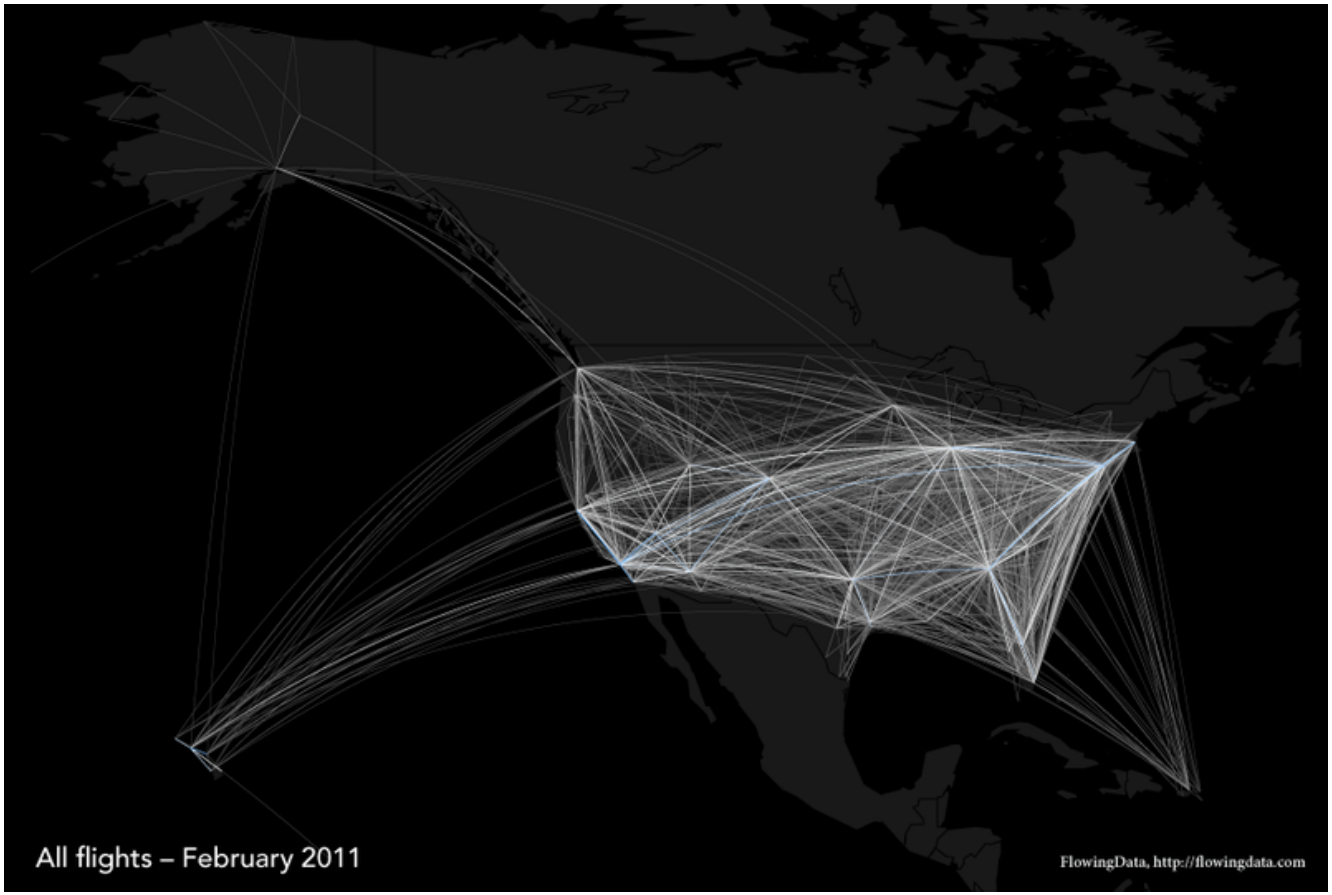
Power laws and fat tails

- The function shown in the previous slide is a so-called power law.
- It states that the distribution of x
- a) is not random,
- b) it has ' order ' ,
- c) high magnitudes of x are not as improbable (as in a random distribution)

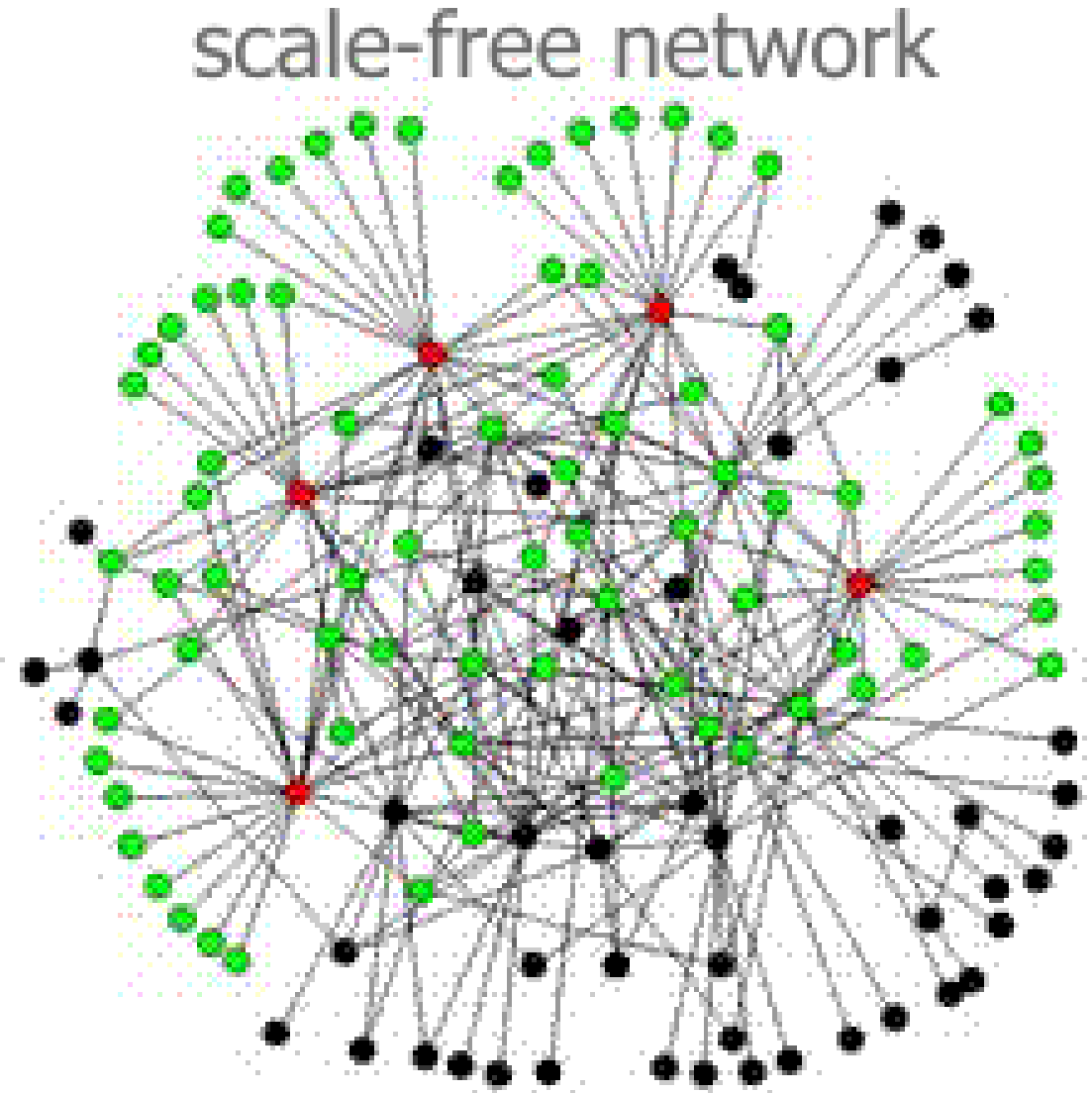
Network theory and graphs

- It has been found and robust evidence has been produced that these power-law phenomena are due to agents' interaction and are thus grounded on connectivity
- The question to be asked: how can connectivity be studied?
- We do know the importance of trade, money, clusters of districts, cities.
- We know the crucial role of information and of its flows across communities and firms.
- The theory of networks sheds some important light on these issues.

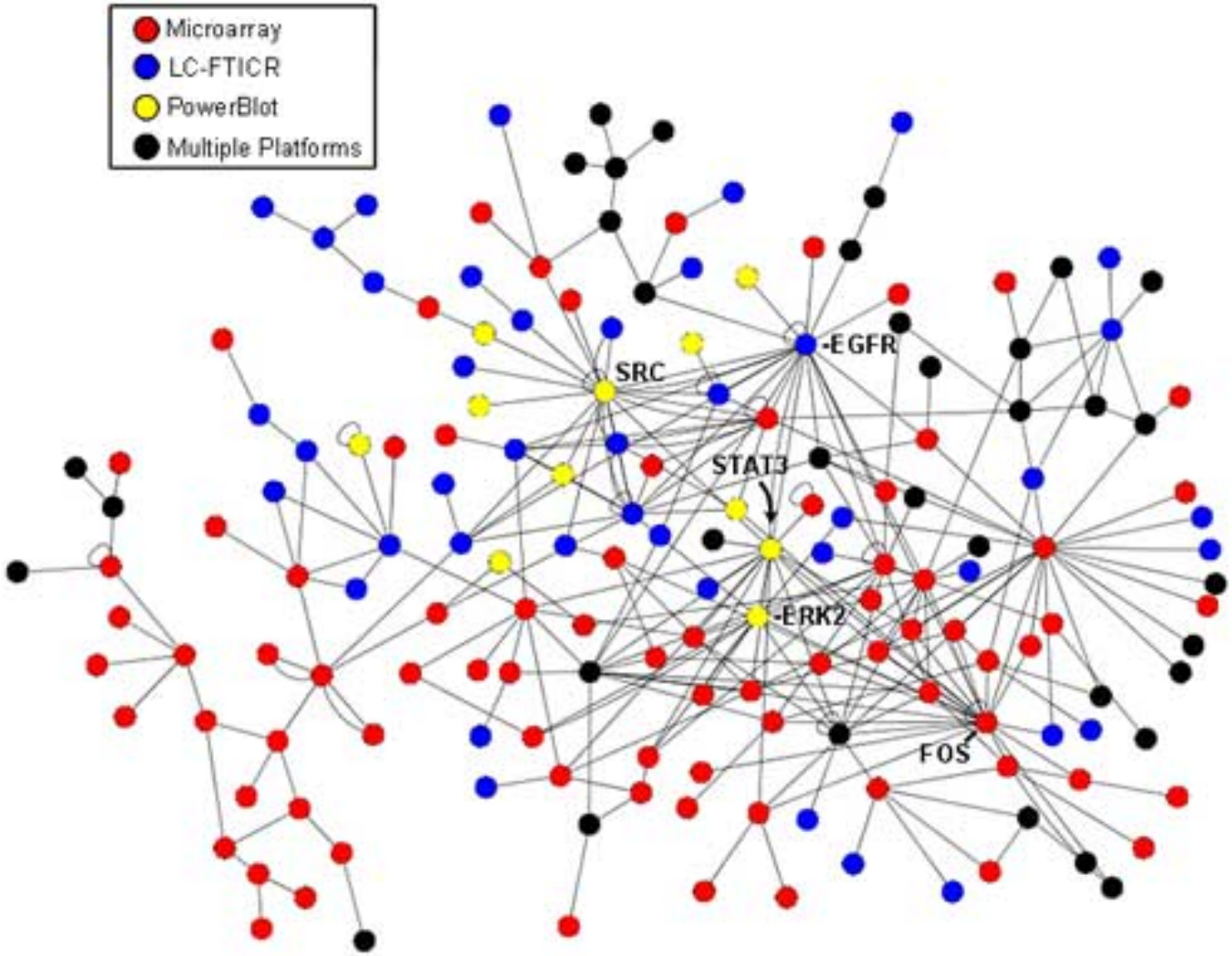
Examples of graphs 1: USA airlines



Examples of graphs 2: Social relations graph



Examples of graphs: a biological network



A benchmark case: random graphs

- It is expedient to start with some definitions.: a graph, define it as composed by a set of nodes, P, and edges, E: $G = G(P, E)$
- A random graph G is a graph of $|P| = N$ nodes connected by$|E| = n$ edges chosen randomly from all possible edges. The order of magnitude of the latter is:

$$n_{max} = C_{N.2} = \frac{N(N-1)}{2}$$

Random graphs of n edges

- Another, interesting order of magnitude: since one can randomly generate a graph by connecting nodes by drawing n edges from a pool of $C_{N.2}$, there are as many as $C_{N.2}$ edges to generate such a graph. This means that one can generate as many as

$$C_{C_{N.2},n} = \frac{\frac{N(N-1)}{2}!}{n!\left[\frac{N(N-1)}{2} - n\right]!}$$

Links

- An example: for instance, if $N = 6$ and $n = 3$; $C_{6,.2} = 15$; $C_{15,3} = 455$ graphs can be generated.
- The interesting question: how does a random graph come into being? More specifically, why is it that certain cliques or clusters come to be when randomness prevails?
- Let us introduce the following idea: nodes are initially entirely unconnected but then proceed to connect them with some probability p. Thus p is the probability that any two nodes be connected.
- The expected number of links:

$$E(\#) = \frac{N(N-1)}{2}p$$

Graphs: basic quantities

- The probability of obtaining a graph with n edges:

$$P(G_0) = p^n(1-p)^{\frac{n(n-1)}{2}-n}$$

- (for instance a specific graph from $N = 6$; $n = 3$ and $p = .2$; say G_0=(adf),$P(adf) = .00055$)
- An important quantity: the average degree of a random graph. How many connections, on average, is a node likely to possess in a random graph?
- A node can connect to as many as $N-1$ other nodes with probability p. Hence:

$$< k >= p(N-1)$$

Approximating p

- If n edges have been successfully established, then we can compute the probability of connection of any two nodes: since there are $\frac{N(N-1)}{2}$ possibilities of connecting and n are the actual ones (the favourable cases), then

$$p = p(N, n) = \frac{2n}{N(N-1)} \approx \frac{2n}{N^2}$$

for large N.

Subgraphs of degree k

- The emergence of some properties, i.e. the shape of graph connections: relationships.
- Let us begin by asking this question: given a random graph, what is the expected number of subgraphs made up of, say, k nodes?
- In a graph of N nodes there are $C_{N,k}$ ways to generate graphs of k nodes, i.e. there are

$$C_{N,k} = \frac{N!}{k!(N-k)!}$$

- From the point of view of the exact shape that a graph acquires, especially if the question is the likely relationships, each subgraph can potentially give rise to $k!$ other graphs .

Subgraphs of degree k, continued

- (e.g. take a graph of $N = 6$ (a, b, c, d, e, f) nodes and consider a subgraph of$k = 3$ nodes: (adf). The nodes in this subgraph can also come in the shape of $(afd, daf, dfa, fad, fda)$.
- In actual and practical problems some allowance must be made for the fact that some of these subgraphs have the same relevance and thus the actual number that each graph can really generate is

$$\frac{k!}{a}$$

- . E.g. if only half are really different, divide by 2 $(a = 2)$.

Connected subgraphs

- Now, the next question: what is the expected number of connected subgraphs if the available edges to connect the k nodes is l and the connection probability is p?

$$E(X_k) = C_{N,k}\frac{k!}{a}p^l = \frac{N!}{k!(N-k)!}\frac{k!}{a}p^l = \frac{N!}{(N-k)!}\frac{p^l}{a} \approx \frac{p^l}{a}N^k$$

  for large$N$ and relatively small k .
- An example, $N = 100;\ k = 6;\ l = 3;$
  $a = 1; p = .2 \rightarrow E(X_k) = 10,000.000 * .008 = 80,000$
- The number of possible graphs is very large but the probability of establishing 'relationships' of as many as l links is small.

Important properties

- As it has been seen if in a random graph there happen to be n connections, then the probability that any two nodes be connected can be estimated to be:

$$p = p(N, n) \approx \frac{2n}{N^2}$$

- It has been found that there exists a critical probability $p_c = p_c(N)$ below which, for $p = p(N) < p_c)$, almost no 'property' of subgraph connections appear.
- Whilst for $p = p(N) \geq p_c$ most such subgraphs connections do!

The critical probability

- To see what such a critical probability is , consider $E(X_k) = (p^l/a)N^k$, the expected number of subgraphs composed of k nodes linked by l edges.
- If p is very, very small, $E(X)$ is likely to be very small too, almost insignificant. At which level of p $E(X)$ becomes surely significant? Consider the following critical probability:

$$p_c(N) = cN^{-\frac{k}{l}}$$

- where c is an arbitrary constant. It follows that

$$E(X_k) = \frac{c^l}{a}$$

Trees, triangles, cycles and complete graphs

- Some cases: the critical probability at which almost every graph contains a subgraph with k nodes and l edges
- - a tree of order k and $(l = k-1)$: $p_c(N) = cN^{-(k/(k-1))}$
- - a cycle or order k and $(l = k)$: $p_c(N) = cN^{-1}$
- a complete subgraph of order k and $(l = \frac{k(k-1}{2}$) : $p_c(N) = cN^{-\frac{2}{k-1}}$

An example

- The critical probability for a graph to contain a completely connected subgraph :
- $k = 10$, hence with $\frac{k(k-1)}{2} = 45$ connections and $c = 2$, is
- $p_c = 43\%$, if $N = 1000$,
- $p_c = 26\%$, if $N = 10000$,
- $p_c = 15,5\%$, if $N = 100000$; thus in graphs of many nodes, cliques start appearing even for low probabilities of setting up a connection.

The degree distribution of random graphs

- Question: what is the probability that a node named k_i has k degrees (connected to k other nodes)? Answer:

$$P(k_i = k) = P(k) = C_{N-1,k} p^k (1-p)^{(N-1)-k}$$

- The expected number of so connected nodes is:

$$E(x_k) = NP(k)$$

The Poisson distribution

- Note that, since $< k > \backsimeq pN$, and for $N \to \infty$;

$$P(k) = e^{-<k>} \frac{< k >^k}{k!}$$

- namely, a Poisson distribution.

The average path length of random graphs.

- The average path length, l_{rand}: average distance between any pair of nodes.
- If on average the path length is l_{rand} , then multiplying $< k >$, the average degree, a number l_{rand} of times, one counts (approximately) all the nodes in the network N.
- Hence $< k >^{l_{rand}} = N$, from which

$$l_{rand} = \frac{logN}{log < k >}$$

- Thus, the average path length scales with the log of the network size.

The clustering coefficient

- The clustering coefficient of a node i having k_i connections can be defined as the number of shared connections of i's neighbours over the total number of connections that i can have:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}$$

- Since the expected number of i's connections is $p\frac{k_i(k_i-1)}{2}$, it follows that $C_i = p$. Since $p \approx \frac{2n}{N^2}$ and $< k >= \frac{2n}{N}$, it is

$$C_i = \frac{< k >}{N}$$

- It is clear that the average clustering coefficient is also $C_{rand} = p$.

A pristine world

- A thought exercise: think of a world with an economic structure hallmarked by primitive means of production and transportation:
- Simple economic units such as have been described in the first lecture, with some trade taking place.
- These economies are likely to be either fully isolated or to entertain trade links, share information, experience and learning, pass on or imitate techniques of production only within a short distance.

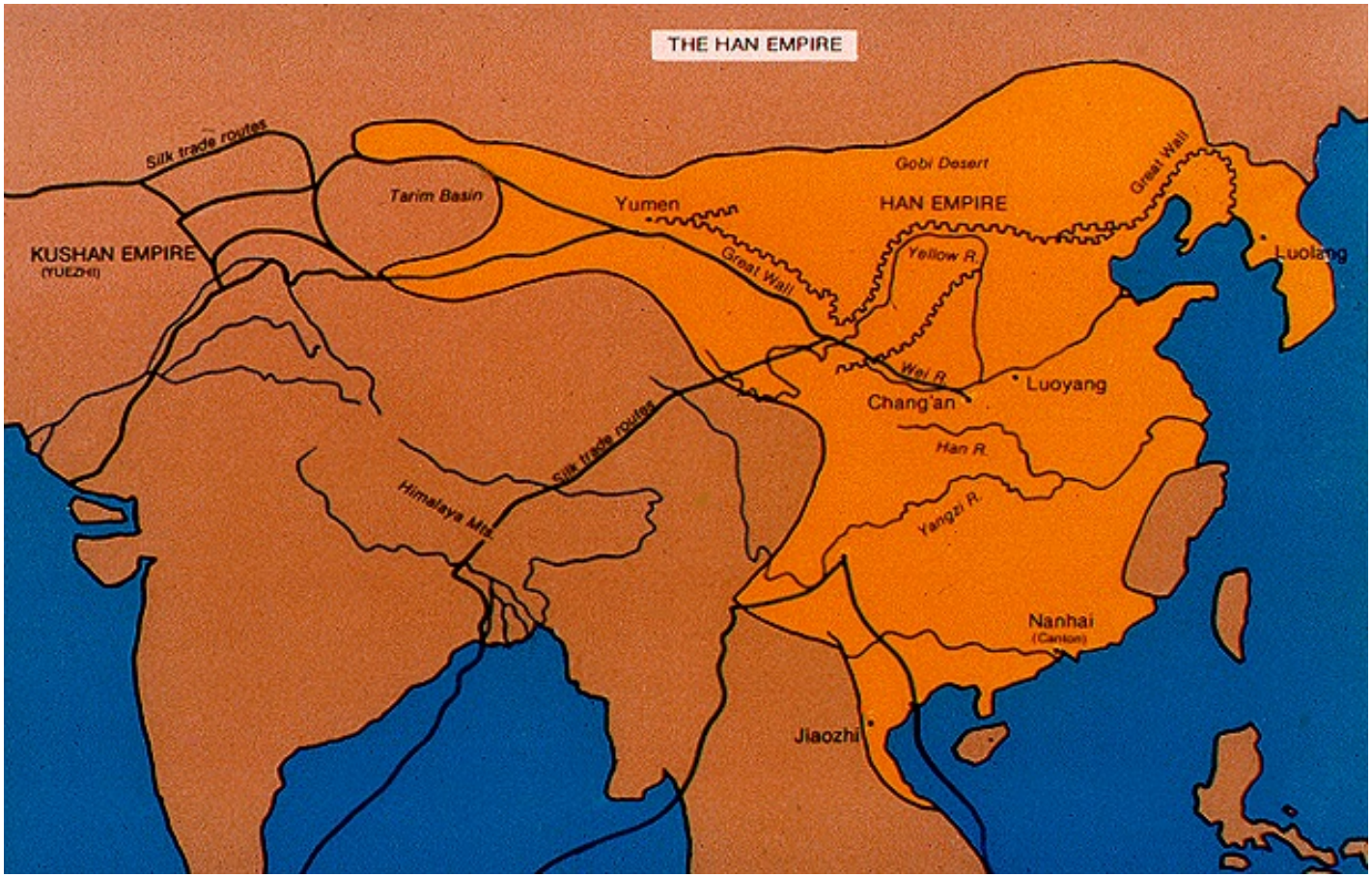
Ancient societies

- Ancient societies roughly resemble this description. But.....
- Vast empires sprang up: think of the Roman and Chinese ones or of the Inca Empire
- Of course, these were sophisticated social set-ups with complex social arrangements, comparatively technologically advanced.
- They had developed roads, bridges, communication linkages.

Rome in the II century C.E.

Geographical Tour: A Tour of Empire

China in the Han dynasty (206 B.C.-220 A.D.)



The Inca empire

An ordered network

- What kept them economically and culturally together? Let us resort to an abstract model.
- Assume an economic society in which nodes cluster, i.e. each node has links with few neighbours at a short distance.
- Such a network can easily be represented by a ring lattice in which each node is immediately connected right and left with a given number of other nodes.
- Assume that this number be $< k >$: the average network degree since all nodes have the same number of links.

Relevant magnitudes

- The relevant magnitudes for an ordered network $N \ggg < k > \ggg logN \ggg 1$;
- The average path length: $l_{ordered} = \frac{N}{2<k>}$
- The clustering coefficient $C_{ordered} = \frac{2E}{<k>(<k>-1)} = \frac{3}{4}\frac{<k>-2}{<k>-1}$, note that the latter is $\frac{3}{4}$ for large networks.
- Since the highest the clustering coefficient can get is 1, $\frac{3}{4}$ is a high clustering coefficient.
- Thus, these networks have a very high clustering coefficient and a very long average path length.

Comparing random with ordered networks

- It is useful to remember that a random network is scarcely clustered, $C_{rand} = p$, and that its path length is rather short, $l_{rand} = \frac{logN}{log<k>}$, it scales with the logarithm of N.
- Should we conclude that a short average path length is always associated with a scarcely clustered network? A long path length with a highly clustered one?
- Our pristine world is one of very long path lengths and is highly clustered: from an economic, social and political point of view very difficult to control and monitor.
- Can intense trade and exchange relationships arise? Can development take place?

The Watts and Strogatz conundrum

- One might be led to think that in order to weave such a society into a manageable social and economic body, plenty of links across the ordered network (the circle) should be implemented.
- The answer is NO! Consider the following procedure: randomly rewire with probability ρ , barring duplications and self wiring.
- This means taking an edge at random away from one node and reconnecting it to another node at random.
- Since the total number of rewirable edges is $N < k > /2$, i.e. the nodes times neighbours on either side, this procedure allows for a long range rewiring of $\rho \frac{N<k>}{2}$ edges.

The network is modified

- The question is: how does the network change as ρ is allowed to change from 0 to 1? Note that in this last case the network becomes entirely random.
- $C(\rho)$ and $l(\rho)$ are expected to vary as a function of ρ, from $C_{ordered} \equiv C(0) \simeq \frac{3}{4}; l_{ordered} \equiv l(0) = \frac{N}{2<k>}$ to $C_{rand} \equiv C(1) = p;\ l_{rand} \equiv l(1) = \frac{logN}{log<k>}$.
- Watts and Strogatz have shown that $l(\rho)$ drops very rapidly with small increases in ρ while $C(\rho)$ varies little with ρ. It follows that there is a large interval in which $l(\rho)$ is short and $C(\rho)$ is high.

A small world

- To see why this happens consider that for small ρ, the path length scales with the system size whilst the clustering coefficient remains roughly constant ($\backsimeq 3/4$).
- As the network becomes more and more random the path length begins to scale logarithmically (small changes) while $C(\rho)$ begins to approach the value of $\rho = \frac{<k>}{N}$.
- It is intuitively clear that the observed 'phenomenon' on the average path length,l, depends on the system size which can here be defined by $< k > N$ on which the probability of rewiring operates: $\rho < k > N$.
- The world is a *small world* .

Relevant quantities

- The actual mathematical form has largely been left to numerical simulations. In any case, approximations indicate that:

$$l(\rho, N, < k >) \backsim \frac{N^{(1/d)}}{< k >) f(\rho, < k >, N)}$$

- The equivalent expression for C is more elaborate and it goes:

$$C(\rho) = \frac{3 < k > (< k > -1)}{2 < k > (2 < k > -1) + 8\rho < k >^2 + 4\rho^2 < k >^2}$$

- The degree distribution: it is very similar to the random graph distribution with a peak $< k >$.

Many (most?) networks are not random

- The problem with the Watts and Strogatz' model is that it applies to networks that have approximately a Poisson distribution so that the most likely number of connections for any given node is just the average $< k >$: from the point of view of connectedness they are about the same.
- Most relevant networks do not have this structure. As noted above, empirical findings have shown that, quite frequently, the distribution of nodes takes the form:

$$P(k) = ak^{-\gamma}$$

$$\pi(k) = a - k\gamma$$

the latter being the log-log form.

The meaning of this distribution

- Since, $P(k) \to 0$ only for $k \to \infty$, it is a distribution that exhibits values significantly different from zero even for very large k's.
- In other words, in such networks there are likely to be few nodes with high k's, some with a sizable k, very many with a small k, i.e. all scales of k are likely to be present.
- The average $< k >$ is not at all representative of the network scale and the ratio of the mean to the variance tends zero.
- This type of networks are often called scale-free networks.

The question

- The question arises: why is it that many networks have such a structure?
- A likely answer is. because their evolution has been such that although they are the result of a stochastic process they do not feature randomness but order.
- Hence, the analytical task is to find a procedure that leads to this result.
- The following is the model conjectured by Albert and Barabasi to deal with this issue.

The Barabasi caper

- These authors have exploited two major and historically well established ideas:
- a) networks grow, i.e. their size N increases with time;
- b) attachment of newly born nodes to existing ones is preferential, i.e. they attach to nodes that already have, in a relative sense, many attachments.
- Proceed as follows:
- a.1) start with a very small number m_0 of nodes and at every time step add a new node with m edges.
- b.1) the probability that the new node attaches to node i depends onk_i, that is it depends on its degree.

Probability and some dynamics

- Hence, the probability that the new node attaches to node i is:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

- Thus, the growth of any node's degree is

$$\frac{\partial k_i}{\partial t} = m\Pi(k_i) = m\frac{k_i}{\sum_j k_j}$$

- but note that, by the above assumptions, $\sum_j k_j = 2mt$, for $m_0$vanishing small.

The solution

- The differential equation is:

$$\frac{\partial k_i}{\partial t} = m\frac{k_i}{\sum_j k_j} = \frac{k_i}{2t}$$

- Solving this differential equation by integration and by assuming that the initial condition for every node is m at some t_i, namely $k(t_i) = m$:

$$k_i(t) = m(\frac{t}{t_i})^{\frac{1}{2}}$$

- all nodes basically evolve in the same way.

Groping towards the nodes degree distribution

- The problem is now to derive the nodes' distribution, given that they do grow in the same way but that their initial condition t_i is different.
- Ask the question: what is the probability that $k_i(t) < k$, that is: $P(k_i(t) < k)$?
- Since the degree depends on the initial condition, i.e. when the node appeared in the network and started to attach: t_i, this question can be rephrased as:

$$P(k_i(t) < k) = P(t_i > \frac{m^2 t}{k^2})$$

- This is equivalent to asking: what is the probability that at time $\bar{t} = \frac{m^2 t}{k^2}$ node i has not yet appeared ? This is a very convenient question.

The probability of a node appearing on the scene

- It is very convenient since we can then ask the question: what is the probability of i being in the network at time t? Since at each point in time a node is added to the network, it is:

$$P_i(t) = \frac{1}{m_0 + t} = P(t)$$

  the same for all i's .
- Thus, the probability that after $\bar{t} = \frac{m^2 t}{k^2}$ periods any node be in the network is $\frac{\bar{t}}{m_0+t} = \frac{1}{m_0+t}\frac{m^2 t}{k^2}$ and that it be not is:

$$P(t_i > \frac{m^2 t}{k^2}) = 1 - \frac{1}{m_0 + t}\frac{m^2 t}{k^2}$$

The frequency distribution

- By differentiating the above :

$$\frac{\partial P(k_i(t) < k)}{\partial k} = \frac{\partial P(t_i > \frac{m^2 t}{k^2})}{\partial k} = P(k)$$

i.e.

$$P(k) = \frac{2m^2 t}{m_0 + t} \frac{1}{k^3}$$

- Asymptotically, for $t \to \infty$

$$P(k) = 2m^2 k^{-3}$$

A power law, scale free distribution

- Setting $2m^2 = \alpha$ and $\gamma = 3$, it is in general:

$$P(k) = \alpha k^{-\gamma}$$

- What has been derived through the above outlined procedure is the observed family of functions $f(x) \sim x^{-\gamma}$.

Conclusion

- In these type of networks, there are likely to be few nodes, that is nodes whose frequency is very small, to which most other nodes are attached
- These nodes 'rule the rooster'; they are the ones who are the source, both quantitatively and qualitatively, of information; through which most energy flows through (electricity grids), that provide most interbank loans (banking networks), that set technological paradigms (user-producer's networks), cities that attract most population (the Zip's law).

Suggested readings

1. Albert R. and Barabasi, A-L. (2002): ' Statistical Mechanics of Complex Networks'. *Review of Modern Physics*. Vol. 74, pp. 47-97
2. Dorogovtsev S.N., Mendes J.F.F. (2003): *'Evolution of Networks: from Biological Nets to the Internet and WWW.'* Oxford University Press. Oxford.
3. Newman M.E.J. (2003): ' The Structure and Function of Complex Networks' *arXiv:cond-mat/0303516v1*.

Further reading:

1. Dorogovtsev S.N., Goltsev A.V., Mendes J.F.F. (2007): 'Critical Phenomena in Complex Networks' *arXiv:070.0010v6[cond-math.stat-mech]*