Introduction to Complex Networks

Marco Tomassini

University of Lausanne
Lausanne, Switzerland
marco.tomassini@unil.ch

http://www.sigevo.org/gecco-2016/

GECCO'16 Companion, July 20-24, 2016, Denver, USA.

http://dx.doi.org/10.1145/2908961.2926985

Contents

What are Complex Networks?

Networks are objects that can be described with the notions of Graph Theory

Roughly speaking, networks are sets of points (vertices) some of which are connected in pairs by links (edges or arcs)

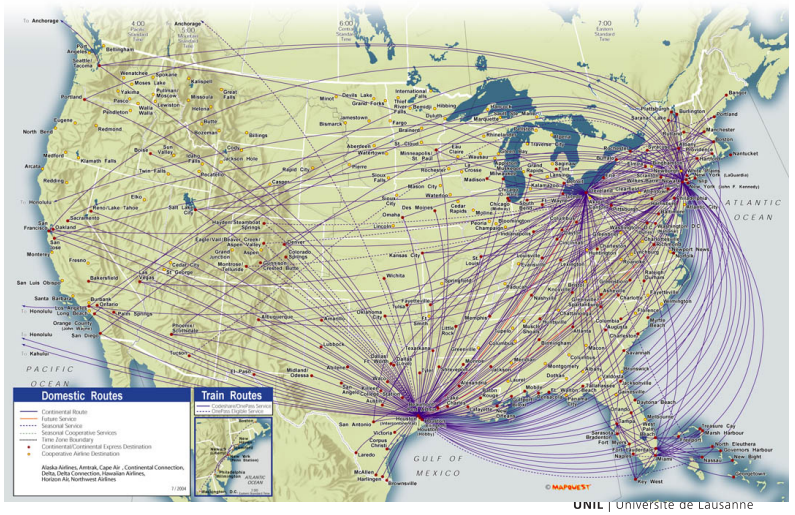
They have been used for the last fifty years at least in the field of social network analysis but were typically very small

They have also been known for a long time in engineering as well, think of the Internet and power networks for instance

They are, explicitly or implicitly, the basic substrate in almost all fields of human and technological activity

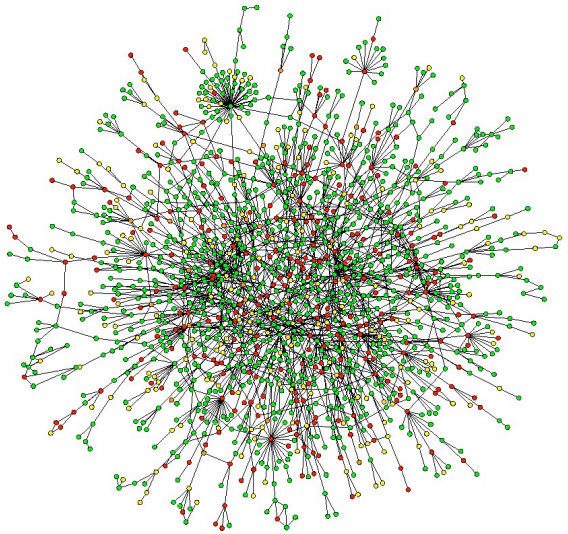
Some Complex Networks 1

The domestic US airlines network



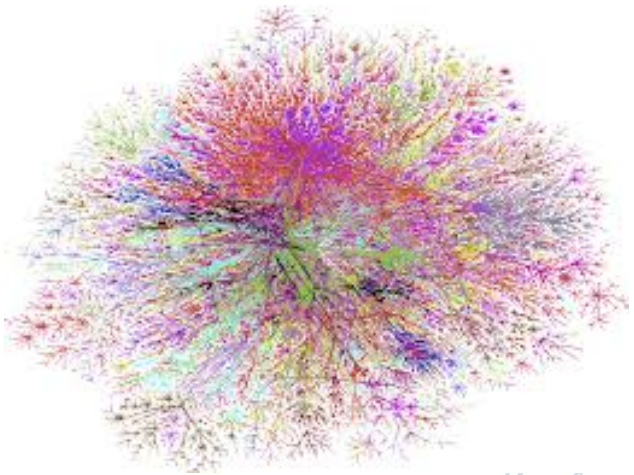
Some Complex Networks 2

Yeast protein interaction network



Some Complex Networks 3

A portion of the Internet network



UNIL | Université de Lausanne

What's New Then?

- In the last fifteen years many large-scale electronic databases of network data have been recorded and analyzed
- New statistical techniques have been used to characterize networks of large size
- Several models have been proposed that approximately reproduce the observed features of large-scale complex networks
- Dynamical processes of many kinds have been modeled and analyzed on networks, e.g. epidemics, search, traffic, diffusion, failures, cascades, growth, games, learning, among others

UNIL | Université de Lausanne

The Main Observations

- Many biological, social, and technological networks appear to have links that are not randomly distributed
- There are short paths between most pairs of nodes
- Nodes are often clustered in groups
- The number of contacts that a node may have can be highly heterogeneous

UNIL | Université de Lausanne

The Elementary Graph Notions Needed 1

The following elementary concepts from graph theory will be needed and are assumed to be common knowledge

- a graph $G(V, E)$, where V is the set of vertices and E is the set of edges or arcs, both finite
- graphs can be undirected (symmetrical relationship), directed, or mixed
- the vertex degree for undirected graphs and the in-degree and out-degree for undirected ones
- a path $p(G)$ as a sequence of vertices or edges and its length $l(p)$
- the adjacency matrix $A(G)$ of a graph G
- weighted graphs in which some numerical value is attached to each edge

UNIL | Université de Lausanne

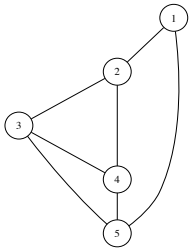
The Elementary Graph Notions Needed 2

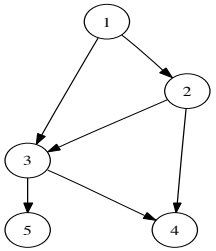
- let $N = |V|$, N is called the order of a graph G. $M = |E|$ is the size of G.
- a complete (undirected) graph has all the possible edges between pairs of vertices: $M = N(N-1)/2 = \binom{N}{2}$.
- A subgraph of G is a subset of a graph's edges and associated vertices that constitutes a graph. That is, $G' = (V', E')$ is a subgraph of $G = (E, V)$ if $V' \subseteq V$ and $E' \subseteq E$.
- A graph is connected if there is a path of finite length between any two vertices. A graph that is not connected consists of a set of connected components.
- The neighborhood $\Gamma(v)$ of a vertex v is the set of vertices that are adjacent to v in G.
- A graph is dense if the number of edges $M \propto N^2$. It is sparse if $M \ll N(N-1)/2$

UNIL | Université de Lausanne

The Elementary Graph Notions Needed 3

Two graphs





And their adjacency matrices:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

UNIL | Université de Lausanne

Some Useful Network Statistics

- Mean degree $\bar{k} = \frac{1}{N}\sum_{i=1}^{N} k_i$
- Degree distribution function $P(k)$: the frequency with which each degree k appears in the graph
- Average path length $\bar{L}$: the mean of all the two-point shortest paths in the graph
- Clustering coefficient C: to what extent my neighbors are neighbors themselves
- Several others including centrality measures, correlations of various kinds etc.

UNIL | Université de Lausanne

Degree Distribution Function (DDF)

The degree distribution $P(k)$ of an undirected graph G is a function that gives the probability that a randomly selected vertex has degree k. $P(k)$ can also be seen as the fraction of vertices in the graph that have degree k.

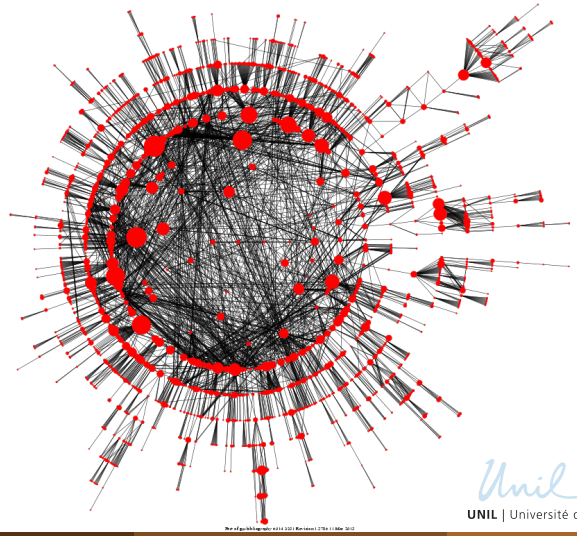
For finite networks the DDF is discrete: $P(k) = \sum_i p_i \delta(k - k_i)$

one often considers the cumulative DDF : $C(k) = \sum_{k=k_{min}}^{k_{max}} P(k)$

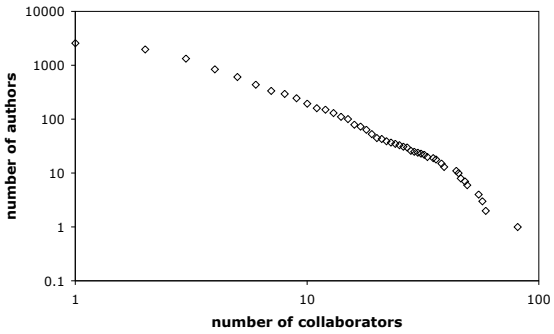
For directed graphs one can consider the outdgree distribution function, and the indegree distribution function.

The GP Coauthorship Network (W. B. Langdon)

The largest connected component of the Genetic Programming collaboration network. Nodes are authors; there is a link between two authors if they have coauthored at least a paper.



The GP Coautorship Network DDF



As in the case of many other complex networks, the DDF is right-skewed (fat-tailed)

Average Path Length

We denote the shortest path between nodes $i, j \in V(G)$ by l_{ij}. The average path length $\bar{L}$ of G is then defined as:

$$\bar{L} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j>i} l_{ij}.$$

The normalizing constant $2/N(N-1)$ is the inverse of the total number of pairs of vertices.

This is a useful measure as it gives an idea of the extension of the network. A complex network that has a short mean path length is said to be a small world

Here "short" means that $\bar{L} \propto log(N)$.
A "long" mean path would instead show a dependence of the type $\bar{L} \propto N^{1/d}$ with a small integer d

Clustering Coefficient

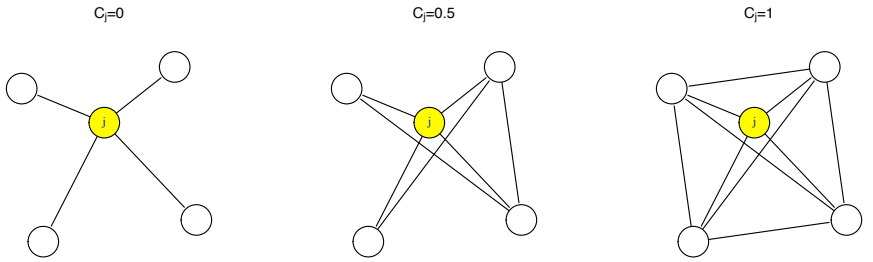
Consider a node j of degree k in a graph. If all k vertices in j's neighborhood were completely connected to each other, forming a clique, then the number of edges would be equal to $\binom{k}{2}$.

The clustering coefficient C_j of node j is defined as the ratio between the e edges that actually exist between the k neighbors and the number of possible edges between these nodes:

$$C_j = \frac{e}{\binom{k}{2}} = \frac{2e}{k(k-1)}. \quad (1)$$

the higher the value of C_j, the more likely it is that two vertices that are adjacent to a third one are also neighbors of each other (transitive closure $\equiv$ formation of triangles)

Example



In the left image the clustering coefficient of node j, $C_j = 0$ since there are no links among j's neighbors.
In the middle image $C_j = 0.5$ because three of the possible maximum six edges among the neighbors of j are present.
In the right image $C_j = 1$ as all the edges that could be there are actually present (it is a clique).

Average Clustering Coefficient

The average clustering coefficient $\bar{C}$ is the average of C_i over all N vertices $i \in V(G)$:

$$\bar{C} = \frac{1}{N} \sum_{i=1}^{N} C_i.$$

The clustering coefficient of a graph G thus expresses the degree of locality of the connections.

Centrality

The concept of centrality is an old one in social network analysis. The idea is to characterize the most central actors in a network, also called sociometric "stars".

One quick measure is the degree of a node: the highest its degree, the most central the node is. But this view is only local.

Instead, in large networks it is also of interest to characterize the nodes or the links that, in some sense, have strategic global significance for the whole network. Some of the measures used for that purpose are:

- node and edge betweennes
- closeness centrality
- eigenvector centrality

Betweennes Centrality

The betweennes b_v of a vertex $v \in V$:

$$b_v = \sum_{i \neq v \neq j} \frac{n_{ij}(v)}{n_{ij}}$$

Where n_{ij} is the total number of shortest paths between i and j, and $n_{ij}(v)$ is the number of those shortest paths that go through v.

Nodes with high betweennes are more central in the sense that they have more control since more traffic goes through them. Nodes with high betweennes play the role of "brokers" in a social sense.

Edge betweennes is defined similarly but for edges instead of nodes.

Correlations and Assortativity

It is of interest to compute correlations of variables represented at the vertices of a complex network.
This can be done through the usual Pearson correlation coefficient or through other similar measures.

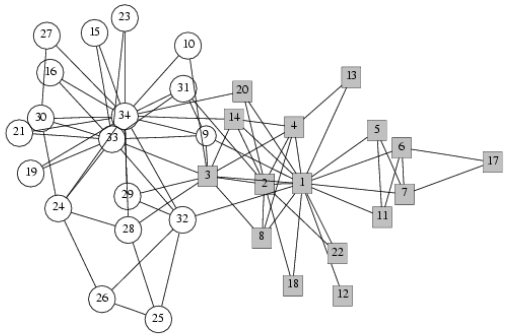
The most immediate one is degree-degree correlation which can be computed through the joint degree distribution function $P(k_1, k_2)$; that is, the probability of finding an edge whose end points have degree k_1 and k_2 respectively.

It has been found that social networks are in general assortative (positive correlation) since vertices of degree k tend to be connected to vertices of similar degree. In contrast, "technological networks" like the Internet are in general disassortative (i.e. node degrees are negatively correlated).

Clusters at the Mesoscopic Scale: Communities

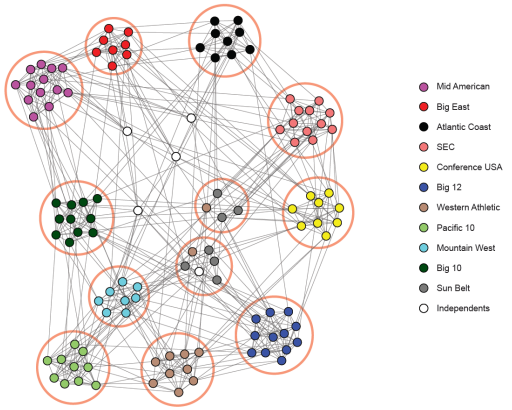
loosely speaking, communities are groups of vertices such that there are many edges inside groups and few to other groups.

The following is a famous example: the so called "Karate club network":



Clusters at the Mesoscopic Scale: Communities

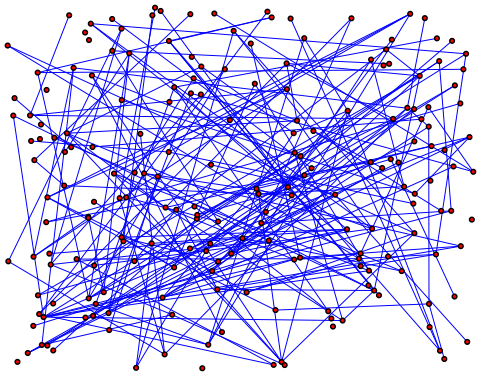
Another example: communities in the NCAA football teams network:



Community detection is a hard computational problem. However, there exist several fast heuristic methods to find clusters, overlapping or not.

Random Graphs 1

Let's have N vertices. Assume that each of the possible $N(N-1)/2$ edges is present with probability p and absent with probability $1-p$.
This gives the $G_{N,p}$ ensemble of equiprobable Random Graphs.
The following is a small random graph with $N = 200$ and $p = 0.01$, which gives a mean degree of about 2



sité de Lausanne

Random Graphs 2

A remarkable property of random graphs is the birth of a giant component when the mean degree reaches 1. After this point, the giant component contains $O(N)$ vertices, while the other smaller components are of size $O(\log N)$.
The phase transition is sharp for $N \to \infty$, it is only approximate in real graphs because of finite-size effects.

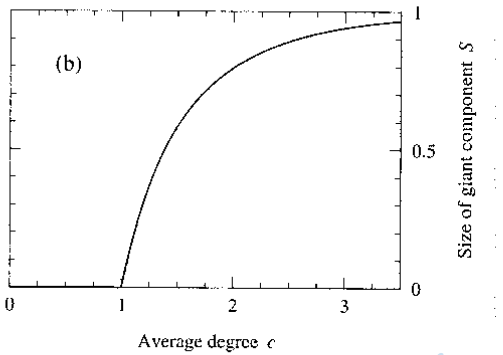
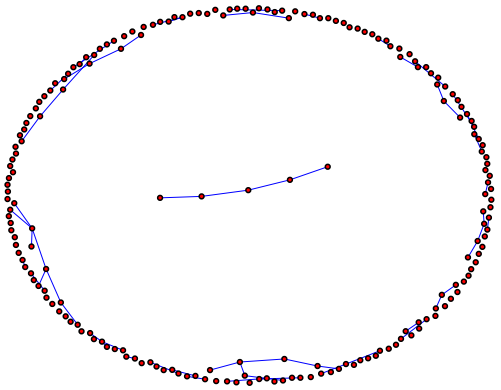


Figure redrawn from M. Newman's book "Networks"

UNIL | Université de Lausanne

Random Graphs 3

Here is a random graph when $p = 0.003$. With $N = 200$, this gives a mean degree $\bar{k}$ of 0.6 which is below the transition critical point of $\bar{k} = 1$



UNIL | Université de Lausanne

Random Graphs 4

For a random graph with connection probability p, the probability $P(k)$ that a random node has degree k is given by

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}.$$

This is the number of ways in which k edges can be selected from a certain node out of the $N-1$ possible edges, given that the edges can be chosen independently of each other and have the same probability p.
Thus, the average degree $\bar{k}$ of a random graph is $(N-1)p \simeq Np$ for large N.

For large N, small p, and constant Np, the binomial distribution can be approximated by a Poissonian one with mean $\bar{k} = Np$:

$$P(k) = e^{-\bar{k}} \frac{\bar{k}^k}{k!}.$$

UNIL | Université de Lausanne

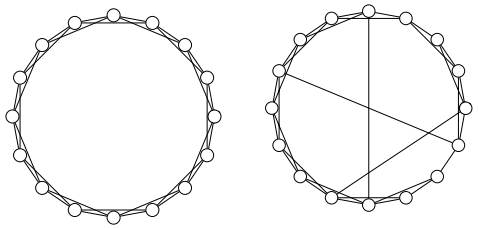
Are Random Graphs Good Models for Complex Networks?

- Are random graphs small words? The answer is yes, as one can show that $\bar{L} = O(\log N)$ for a random graph
- Is the clustering coefficient high? No, the CC is $p \simeq \bar{k}/N$ and it tends to 0 for large N
- Are the node degrees spread over an heterogeneous range? No, because they are strongly peaked around the mean Np

In conclusion, random graphs have the small world property but they are not good models for actual complex network because of their low clustering and their homogeneous degree distribution

Watts and Strogatz Small World Networks 1

The starting point is a lattice in which one goes systematically through each node and, with probability β, its links are rewired toward a randomly chosen node avoiding duplicate edges.

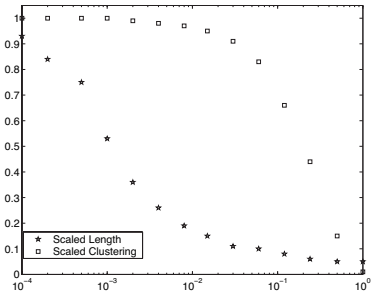


(a) (b)

(a) Regular one-dimensional lattice with $k = 4$. (b) A small-world graph obtained by randomly rewiring some of the nearest-neighbor links.

Watts and Strogatz Small World Networks 2

Lattices have high clustering and high mean path lengths. There exist a range of β for which rewired networks are small-worlds and have a high clustering coefficient. The figure shows scaled clustering coefficient and scaled path length as a function of β



Watts-Strogatz networks still have a peaked degree distribution function.

Barábasi-Albert Model 1

This is a growing network model:

- Start with a small clique of nodes (a fully connected graph) of N_0 nodes and M_0 edges
- At each time step a new node is added and forms m new links to m existing nodes
- The probability $\pi(k_i)$ with which an incoming node forms an edge with an existing node i is:

$$\pi(k_i) = \frac{k_i}{\sum_j k_j},$$

i.e., the higher i's degree is, the larger the probability: this is called preferential attachment

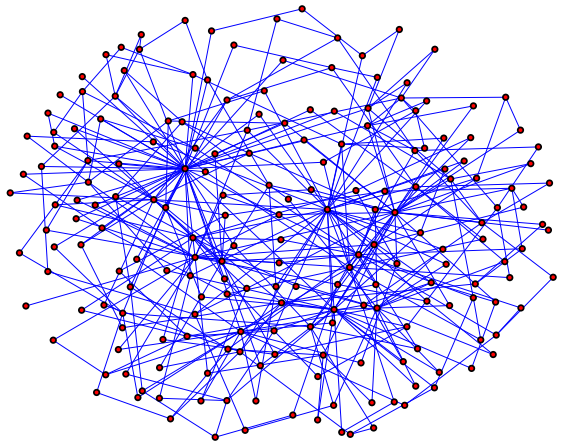
Barábasi-Albert Model 2

- at time step t the graph will have $N_t = N_0 + t$ vertices and $M_t = M_0 + mt$ edges
- the number of nodes with comparatively high degree should intuitively increase with increasing t
- such a growing graph evolves into a stationary scale-free network with a power-law probability distribution of the degree $P(k) \sim k^{-\gamma}$, with $\gamma \sim 3$.

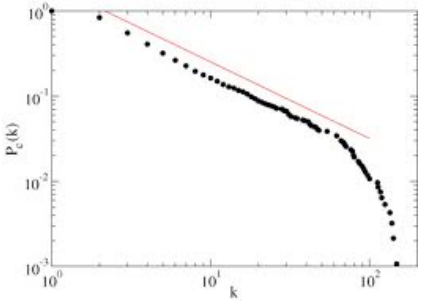
This kind of process imitates, to some extent, the growth of the web, in which already successful pages receive more links; citation networks, in which important papers get cited many times, and many others in which "popular" vertices attract more links when the network forms and evolves

Barábasi-Albert Model 3

An example of a small (N=200, m=2) computer-generated Barabási-Albert network. The presence of a few highly connected nodes (hubs) and of many poorly connected ones is clearly visible



Airline Network: Empirical Degree Distribution



The empirical cumulative degree distribution for the North American part: scale-free with a cutoff. The graph has 932 nodes (airports) and the av. path length is ≈ 4
Other world regions, e.g. Europe, have similar properties

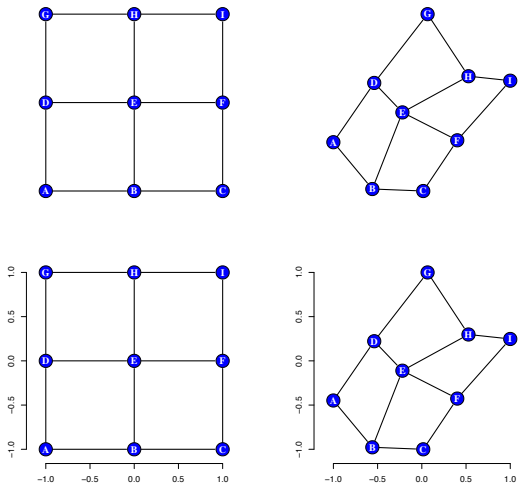
Discussion

- The BA model is closer to many measured complex networks than either RGs or WS networks
- However, the clustering coefficient is too low
- the model has a fixed exponent (≈ 3) of the power-law while a range of exponents between 1.5 and 3.5 have been observed

However:

- It is possible to construct random graphs with arbitrary degree distributions given a degree sequence $\{k_1, k_2, \ldots . k_N\}$
- More recent models exist that closely match the statistical properties of social and other real networks
- It remains nevertheless true than any real network is unique and cannot be exactly generated by any model

Relational Space and Geographical Space



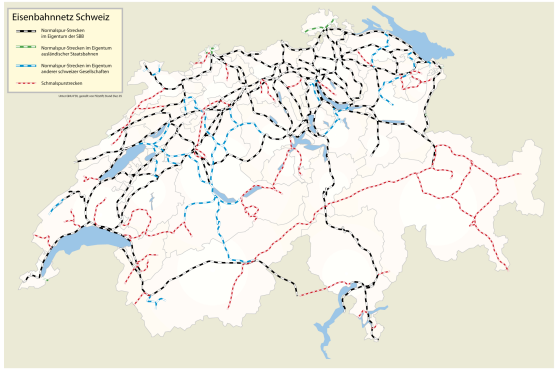
Spatial Networks

Spatial networks are embedded in metric space, usually the ordinary physical two- or three-dimensional space where actual distances become important

There are many networks in society that are of this kind: transportation, communication, biology, roads, streets, power grids, water distribution, ad-hoc networks ...

Some of those networks are also planar: edges do not intersect in the 2-D plane. Examples: road networks (approximately), rail networks.

Railways, Subways



The Swiss rail system: a planar spatial network with $N = 1613$ nodes (stations) and $E = 1680$ edges (connections) between stations.

Swiss Railway

the mean degree $\bar{k} \approx 2.1$

average shortest path $\bar{L} \approx 47$ which is $O(\sqrt{(N)})$ similar to 2-D lattices

the degree distribution $P(k)$ is peaked (exponential) instead of being broadscale

Here we see that geographical and economical constraints play a fundamental role in determining the possible network structure: large degrees are hindered

Results are qualitatively similar for other rail, tramway, and subway systems

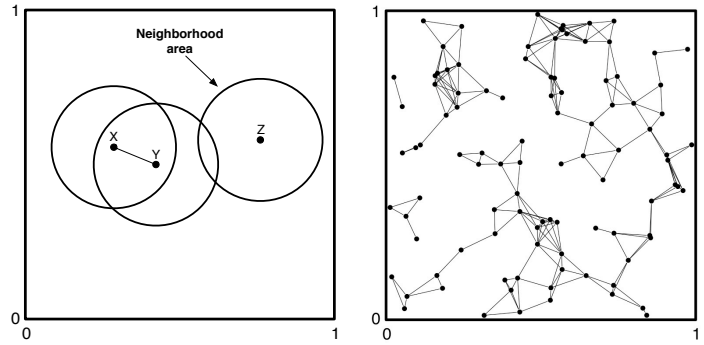
Random Geometric Graphs 1

The Random Geometric Graph (RGG) is a standard spatial network model that plays a role for spatial networks similar to the one played by the Erdös-Rényi random graph for relational ones

- N nodes are placed on the unitary space $\Omega \in \mathbb{R}^d$ with uniform distribution.
- an edge is created for every pair of nodes whose distance is $r < R$.

Random Geometric Graphs 2

here $d = 2$, $N = 100$, and $R = 0.13$



The resulting degree distribution $P(k) \propto 1/k!$ is Poissonian but, contrary to ER random graphs, the clustering coefficient is high

Summing Up

- Spatially embedded networks are very common and important in practice; they have been comparatively less studied than relational networks
- Spatial networks feature natural constraints such as geographic, technological, and economical that strongly influence their structures
- Because of the above most actual spatial networks have a peaked degree distribution function, a relatively long mean path length, and high clustering
- Several useful models of spatial networks do exist: these can be benchmarked and compared to actual empirically measured graphs

Time Evolution of Networks

The networks we have seen up to now have been supposed to be static, which is adequate if they do not change in time or they change so slowly that they can be considered fixed. As well, this description is acceptable if we are interested in a particular "frozen" time snapshot of a given evolving network.

However, it is obvious that many networks do not stay the same as time goes by: traffic networks, crowd networks, ad-hoc mobile networks of communication devices, and many many others such as social networks do change at their own rate.

Network Dynamics

The following processes may contribute to network dynamics, and can make the system an open, non-equilibrium one, where both the number of nodes and links may fluctuate:

- New nodes may join the network
- Nodes may leave the network
- New links can be established among existing nodes
- Links can be cut among existing nodes

In some cases, a first acceptable approximation is to consider the number of vertices N constant and only the number of links M and their endpoints may change (closed system approximation)

An Example of Network Evolution: E-mail Network

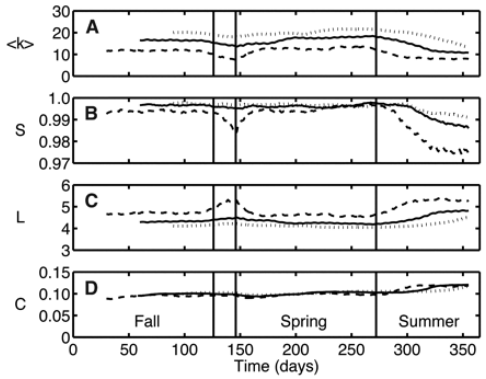
A study comprising 43,553 students, faculty, and staff at a large university, in which interactions between individuals are inferred from time-stamped e-mail headers recorded over one academic year.

Use of e-mail communication to infer the underlying network of social ties is supported by recent studies reporting that use of e-mail in local social circles is strongly correlated with face-to-face and telephone interactions.

The instantaneous network at any point in time includes all pairs of individuals that sent one or more messages in each direction during the past 60 days.

from: G. Kossinets and D. J. Watts, "Empirical Analysis of an Evolving Social Network", Science, 311, 5757, 88-90, 2006.

Some Time-Resolved Statistics

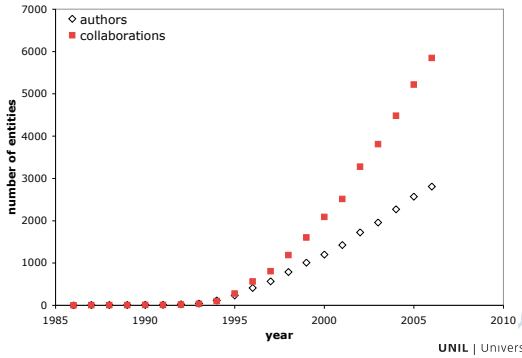


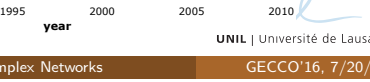
Evolution of mean degree $\bar{k}$, fractional size S of the largest component, mean path length L in the largest component, and clustering coefficient C. Data computed with a smoothing window of 30 days (dashed), 60 says (thick), and 90 days (dots).

Those network properties stay relatively stable, depending of the time-window that is used. Of course larger changes correspond to particular periods like end of semesters and vacations.

Evolution of the GP Coauthorship Network 1986-2006

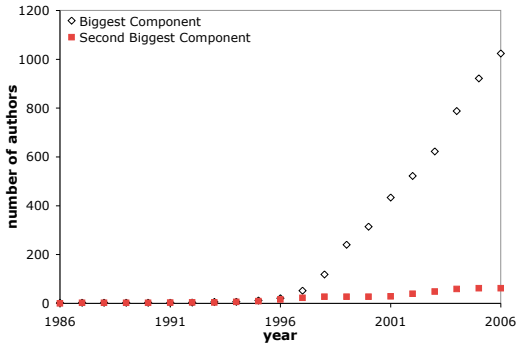
This network evolves more slowly and increases its size and density.
The following figure shows the increase of the number of vertices (collaborators) and of edges (collaborations).
Both grow faster than linear but edges grow faster than nodes because many papers have more than two authors and new authors may collaborate with authors already in the network.





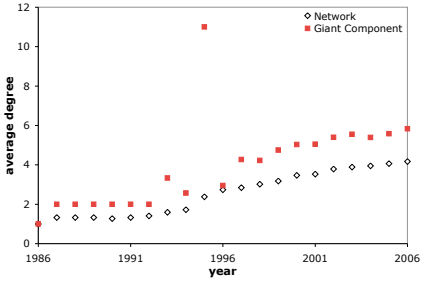
Evolution of the GP Coauthorship Network 1986-2006

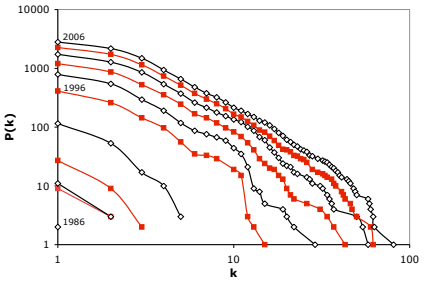
Evolution of the size of the first and second largest components in the GP coauthorship graph. The largest one tends to include more and more nodes as time goes by. This is an instance of a general phenomenon in growing networks.



Evolution of the GP Coauthorship Network 1986-2006

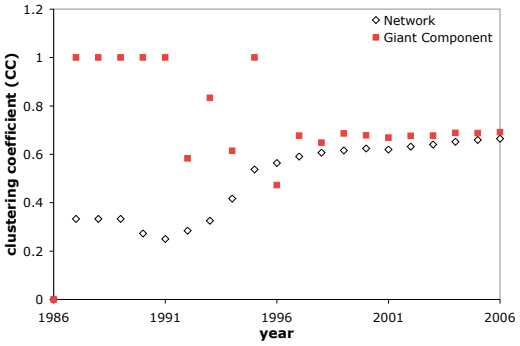
Evolution of the average degree (left), and of the degree distribution (right)





Evolution of the GP Coauthorship Network 1986-2006

Evolution of the clustering coefficient in the whole graph and in the largest component



Contagion: the SI Model and Selection Pressure in EAs

In Evolutionary Algorithms (EAs) the intensity of selection controls the exploitation/exploration tradeoff.

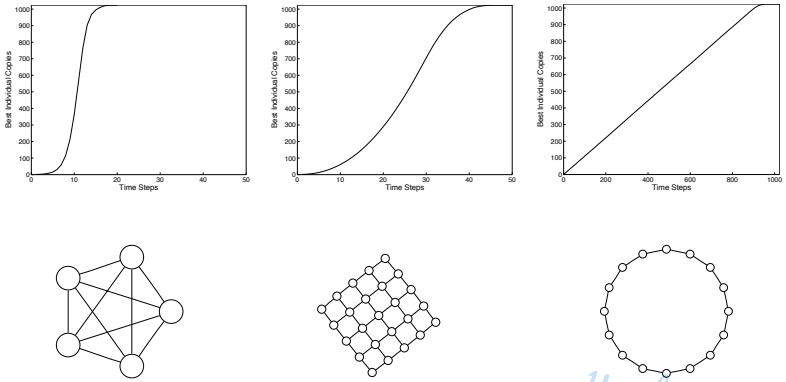
The higher the selection pressure, the more exploitative the EA.

One straightforward method for changing the selection pressure exogenously, i.e. without tampering with selection methods and their parameters, is to keep a standard selection algorithm, say tournament or ranking selection, and vary the population structure seen as a graph. The standard population structure is the well-mixed population, which corresponds to a complete graph.

Selection pressure, in turn, is conventionally measured by its takeover time, i.e. the time it takes for a best individual to take over the entire population under a given selection method.

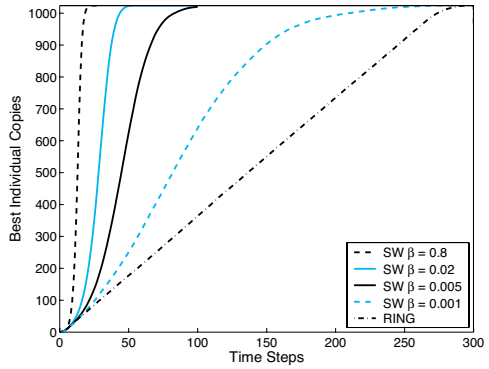
Selection Intensity on Structured Populations

The population structure has a very marked influence on the takeover times and thus on selection intensity (pop. size 1024, binary tournament selection):



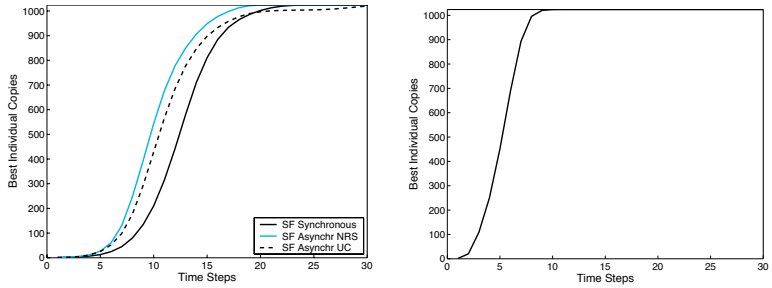
Selection Intensity on Complex Networks: WS Networks

The selection pressure can be varied in a wide range by using a Watts-Strogatz small-world network with different values of the rewiring probability β.



Selection Intensity on Complex Networks: BA Networks

On BA networks takeover times are extremely fast, at least as fast as in well-mixed populations (left image). If the best individual is a hub in the graph the propagation is even faster (right image)



Selection pressure is too strong: not good for EAs

Epidemics: the SIR Model

The previous system resembles what is called the SI (Susceptible-Infected) model in epidemics: An individual is either susceptible or infected. When in contact with an infected individual, a susceptible becomes infected with a certain probability. After a certain time, everybody is infected, which is an absorbing state of the system.

A more epidemiologically realistic model is the SIR (Susceptible-Infected-Recovered) model. The state diagram is:

$$S \rightarrow I \rightarrow R$$

Here "Recovered" may mean either that the individual has got immunity and can no longer catch the infection, or that it is dead (removed).

Epidemics: the SIR Model on Well-Mixed Populations

The mean-field SIR model gives rise to three coupled differential equations for the time evolution of the fractions of susceptible, infected, and recovered individuals. Numerical solution with given initial conditions and actual values for the average rate of encounters and average recovery rate gives the following curves:

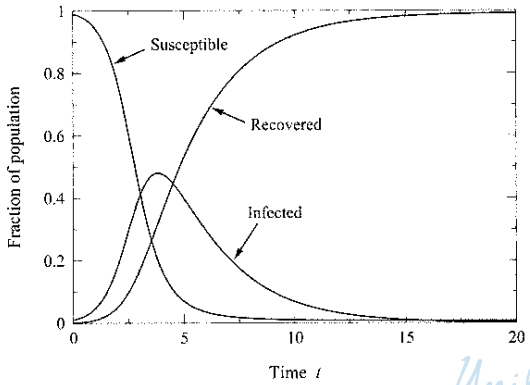


Figure redrawn from M. Newman's book "Networks"

The "Black Death" Propagation

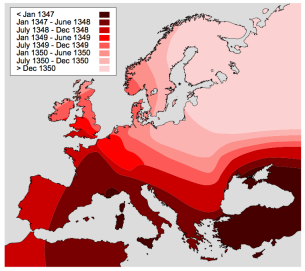
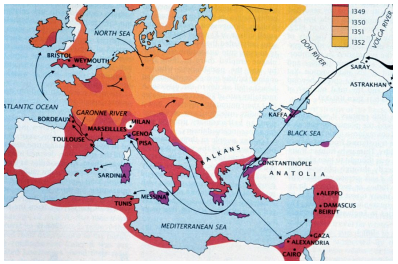
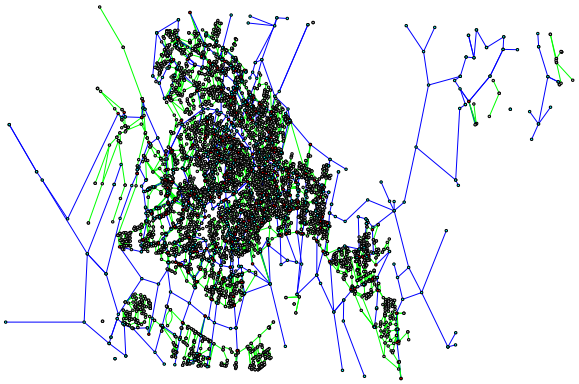


FIG. 1: The spread of the Black Death across Europe in the 14th century, after Sherman and Salisbury [18]. Observe that the disease advanced as a wave of infection across the continent at a more or less constant speed for over three years.



Venice as a Mixed Spatial Network

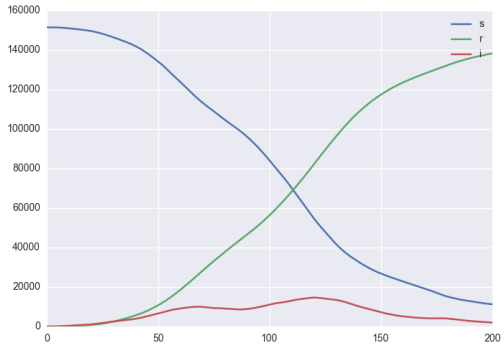
Work by G. Colavizza, PhD Student, EPFL Lausanne



Green edges are land routes; Blue edges are are water routes. Grey vertices are land nodes, red nodes are mixed land and water nodes, and cyan nodes are water nodes only.

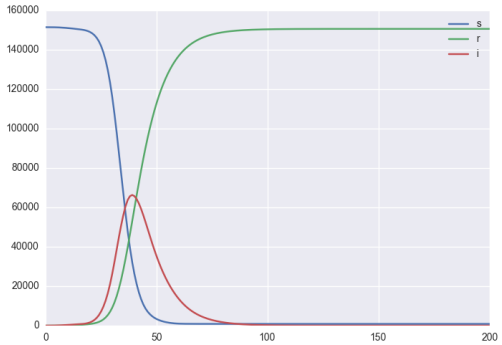
Plague Propagation In Venice: No Naval Travel

Results of a computer simulation of a SIR process where people only move in-land. The network is spatial with a relatively high mean path length (34). This results in a low-infection process.



Plague Propagation Including Naval Travel

The following results include a naval travel probability of only 0.15. This is enough to make the network a rather small-world one thanks to the water travel shortcuts ($\bar{L} = 8.6$). The result is a higher infection rate, more similar to the well-mixed population case.



UNIL | Université de Lausanne

Where to go from here?

An excellent textbook that offers a complete coverage of the field:

M. E. J. Newman, Networks: An Introduction, Oxford University Press, 2010.

An easier to read book with an emphasis on social and economic networks:

D. Easley and J. Kleinberg, Networks, Crowds, and Markets, Cambridge University Press, 2010.

It can also be freely downloaded as a pdf from:

https://www.cs.cornell.edu/home/kleinber/networks-book/

An intermediate level good book:

A. Barrat, M. Barthélemy, A. Vespignani, Dynamical Processes on Complex Networks, Cambridge University Press, 2008.

UNIL | Université de Lausanne