Statistical Mechanics of Networks

Together with

C. Caretta, M. Catanzaro, F. Colaiori, D. Garlaschelli, L. Pietronero, V. Servedio
INFM and Dipartimento di Fisica, Università “La Sapienza” Roma, Italy

L. Laura, S. Leonardi, S. Millozzi, A. Marchetti-Spaccamela
Dipartimento di Sistemistica e Elettronica, Università “La Sapienza” Roma Italy

P. De Los Rios\textsuperscript{a}, G. Bianconi\textsuperscript{b}, A. Capocci\textsuperscript{b}
\textsuperscript{a}Université de Lausanne, \textsuperscript{b}Université de Fribourg, Switzerland

S. Battiston, A. Vespignani
Ecole Normale Supérieure and Université de Paris Sud, Paris France
### Contents

**Part 1  20-11-2003**

**BASICS**
- A. Networks as complex structures (9)
- B. Fractals, Self-similarity (11)
- C. Self-organization (8)
- D. Data of scale-free networks (5)
- E. Basic of Graphs (8)

**Part 2  27-11-2003**

**REAL GRAPHS**
- A. Technological data: Internet (10)
- B. Technological data: WWW (7)
- C. Social data: Finance (13)
- D. Biological data: Food Webs and Proteins

**Part 3  4-12-2003**

**REAL TREES**
- A. Geophysical data: the River Networks
- B. Biological data: Taxonomy and Food Webs
- C. Community Structures

**Part 4  11-12-2003**

**MODELS**
- A. Random Graphs (Erdös-Renyi)
- B. Small world
- C. Preferential attachment
- D. Fitness models
COSIN
COevolution and Self-organisation In dynamical Networks

FET Open scheme RTD Shared Cost Contract IST-2001-33555
http://www.cosin.org

- Nodes: 6 in 5 countries
- Period of Activity: April 2002-April 2005
- Budget: 1.256 M€
- Persons financed: 8-10 researchers
- Human resources: 371.5 Persons/months
The graph size is the number of its vertices.
The graph measure is the number of its edges.
The degree of a vertex in a graph is the number of edges that connects it to other vertices.
In the case of an oriented graph the degree can be distinguished in in-degree and out-degree.
Whenever all the vertices share the same degree the graph is called regular.
A series of consecutive edges forms a path.
  oThe number of edges in a path is called the length of the path.
  oA Hamiltonian path is a path that passes once through all the vertices (not necessarily through all the edges) in the graph.
  oA Hamiltonian cycle is a Hamiltonian path which begins and ends in the same vertex.
  oAn Eulerian path is a path that passes once through all the edges (not necessarily once through all the vertices) in the graph.
  oAn Eulerian cycle is an Eulerian path which begins and ends in the same edge.
Whenever all the vertices share the same degree the graph is called **regular**.

A series of consecutive edges forms a **path**.

- The number of edges in a path is called the **length** of the path.
- A Hamiltonian path is a path that passes once through all the vertices (not necessarily through all the edges) in the graph.
- A Hamiltonian cycle is a Hamiltonian path which begins and ends in the same vertex.
- An Eulerian path is a path that passes once through all the edges (not necessarily once through all the vertices) in the graph.
- An Eulerian cycle is an Eulerian path which begins and ends in the same edge.
A graph is **connected** if a path exists for any couple of vertices in the graph.

A graph with no cycles is a **forest**. A **tree** is a connected forest.

The **distance** between two vertices is the shortest number of edges one needs to travel to get from one vertex to the other.

Therefore the **neighbours** of a vertex are all the vertices which are connected to that vertex by a single edge.

A **dominating set** for a graph is a set of vertices whose neighbours, along with themselves, constitute all the vertices in the graph.

A graph with size \( n \) cannot have a measure larger than \( m_{\text{max}} = n(n-1)/2 \). When all these possible edges are present the graph is **complete** and it is indicated with the symbol \( K_n \).

The opposite case happens when there are no edges at all. The measure is 0 and the graph is then **empty** and it is indicated by the symbol \( E_n \).

The **diameter** \( D \) of a graph is the longest distance you can find between two vertices in the graph.

A complete **bipartite clique** \( K_{i,j} \) is a graph where every one of \( i \) nodes has an edge directed to each of the \( j \) nodes.

The **clustering coefficient** \( C \) is a rougher characterization of clustering with respect to the clique distribution. \( C \) is given by the average fraction of pair of neighbours of a node that are also neighbours each other. For an empty graph \( E_n \) \( C=0 \) everywhere. For a complete graph \( K_n \), \( C=1 \) everywhere.

A **bipartite core** \( C_{i,j} \) is a graph on \( i+j \) nodes that contains at least one \( K_{i,j} \) as a subgraph.
Router connections at small level produce a complex Internet structure.
Previous maps have been computed through extensive collection of traceroutes.

gcalda@pil.phys.uniroma1.it> traceroute www.louvre.fr

1 141.108.1.115 Rome pcpil
2 141.108.5.4 Unknown
3 193.206.131.13 Unknown rc-infnrmi.rm.garr.net
4 193.206.134.161 Unknown rt-rc-1.rm.garr.net
5 193.206.134.17 Unknown mi-rm-1.garr.net
6 212.1.196.25 South Cambridgesh garr.it.ten-155.net
7 212.1.192.37 South Cambridgesh ch-it.ch.ten-155.net
8 212.1.194.14 Genève geneva5.ch.eqip.net
9 195.206.65.105 Genève geneva1.ch.eqip.net
10 0.0.0.0 Unknown No Response
11 193.251.150.30 Unknown p6.genar2.geneva.opentransit.net
12 193.251.154.97 PARIS, FR p43.bagbb1.paris.opentransit.net
Results are that we can quantify the hierarchical nature of the AS connections

\[ P(A) \propto A^{-2} \]

Plot of the \( C(A) \) show the same optimisation of the Food webs

\[ C(A) \propto A \]
skitter is a tool for actively probing the Internet in order to analyze topology and performance.

• **Measure Forward IP Paths**
  skitter records each hop from a source to many destinations. by incrementing the "time to live" (TTL) of each IP packet header and recording replies from each router (or hop) leading to the destination host.

• **Measure Round Trip Time**
  skitter collects round trip time (RTT) along with path (hop) data. skitter uses ICMP echo requests as probes to a list of IP destinations.

• **Track Persistent Routing Changes**
  skitter data can provide indications of low-frequency persistent routing changes. Correlations between RTT and time of day may reveal a change in either forward or reverse path routing.

• **Visualize Network Connectivity**
  By probing the paths to many destinations IP addresses spread throughout the IPv4 address space, skitter data can be used to visualize the directed graph from a source to much of the Internet.

http://www.caida.org/tools/measurements/skitter
2A Internet

EuroRings™

Troisieme Cycle Suisse Romande
Stat. Mech. of Networks-
This happens at both domain and router server

$P(k) =$ probability that a node has $k$ links

Faloutsos et al. (1999)
Internet maps measurements
- CAIDA
- NLANR
- Mercator project
- IPM
- Bell lab.s
Troisieme Cycle Suisse Romande
Stat. Mech. of Networks-

Vazquez Pastor-Satorras and Vespignani
PRE 65 066130 (2002)

TABLE I. Total number of new ($N_{\text{new}}$) and deleted ($N_{\text{del}}$) nodes in the years 1997, 1998, and 1999. We also report the number of deleted nodes with connectivity $k > 10$.

<table>
<thead>
<tr>
<th>Year</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{\text{new}}$</td>
<td>309</td>
<td>1990</td>
<td>3410</td>
</tr>
<tr>
<td>$N_{\text{del}}$</td>
<td>129</td>
<td>887</td>
<td>1713</td>
</tr>
<tr>
<td>$N_{\text{del}}(k &gt; 10)$</td>
<td>0</td>
<td>14</td>
<td>68</td>
</tr>
</tbody>
</table>

TABLE II. Average properties of the Internet for three different years. $N$, number of nodes; $E$, number of connections; $\langle k \rangle$, average connectivity; $\langle c \rangle$, average clustering coefficient; $\langle r \rangle$, average path length; $\langle b \rangle$, average betweenness. Figures in parentheses indicate the statistical uncertainty from averaging the values of the corresponding months in each year.

<table>
<thead>
<tr>
<th>Year</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>3112</td>
<td>3834</td>
<td>5287</td>
</tr>
<tr>
<td>$E$</td>
<td>5450</td>
<td>6990</td>
<td>10100</td>
</tr>
<tr>
<td>$\langle k \rangle$</td>
<td>3.5(1)</td>
<td>3.6(1)</td>
<td>3.8(1)</td>
</tr>
<tr>
<td>$\langle c \rangle$</td>
<td>0.18(3)</td>
<td>0.21(3)</td>
<td>0.24(3)</td>
</tr>
<tr>
<td>$\langle r \rangle$</td>
<td>3.8(1)</td>
<td>3.8(1)</td>
<td>3.7(1)</td>
</tr>
<tr>
<td>$\langle b \rangle/N$</td>
<td>2.4(1)</td>
<td>2.3(1)</td>
<td>2.2(1)</td>
</tr>
</tbody>
</table>
 Troisieme Cycle Suisse Romande  
 Stat. Mech. of Networks-
Nodes: (static) HTML pages

Edges (directed): hyperlinks between pages
Why are we interested in the WebGraph?

From link analysis:
- Data mining (ex: PageRank)
- Sociology of content creation
- Detection of communities

With a “good” WebGraph model:
- Prove formal properties of algorithms
- Detect peculiar region of the WebGraph
- Predict evolution of new phenomena
Models for the WebGraph:

• Random Graph (Erdös, Renyi)
• Evolving networks (Albert, Barabasi, Jeong)
• “Copying” models (Kumar, Raghavan,…)
• ACL for massive graph (Aiello, Chung, Lu)
• Small World (Watts, Strogats)
• Fitness (Caldarelli, Capocci, De Los Rios, Munoz)
• Multi-Layer (Caldarelli, De Los Rios, Laura, Leonardi)
In–degree density distribution

Out–degree density distribution

Albert Barabasi *Emergence of scaling in random networks*
Kumar et al., *Stochastic models for the WebGraph*
Broder et al., *Graph structure in the web*
• Bow-tie structure
• Small World for the SCC and the weakly connected components

Broder et al., *Graph structure in the web*
Cyber Communities

- **Explicit** (or “self-aware”) communities:
  1. Webrings
  2. Newsgroup users
  3. Gnutella, Morpheus, etc. users

- **Implicit** communities:
  1. Fan-Center Bipartite Cores

---

Kumar et al., *Crawling the Web for Emerging Cyber Communities*
Fractal properties

- **TUC** - Thematically Unified Cluster, for example:
  1. By content
  2. By location
  3. By geographical location
  …and…
  4. Random collection of websites
  5. Hostgraph

Dill et al., *Self-similarity in the web*
Probably the most complex system is human behaviour! Even by considering only the trading between individuals, situation seem to be incredibly complicated.

Econophysics tries to understand the basic “active ingredients” at the basis of some peculiar behaviours. For example price statistical properties can be described through a simple model of agents trading the same stock.

“A Prototype Model of Stock Exchange”
Some of the phenomena in finance can be described by means of graphs

- Stock price correlations
  - J.-P. Onnela, A. Chackraborti, K. Kaski, J. Kertész, A. Kanto
  - G. Bonanno, G. Caldarelli, F. Lillo and R. N. Mantegna
    http://xxx.lanl.gov/abs/cond-mat/0211546

- Portfolio composition
  - D. Garlaschelli, S. Battiston, M. Castri, V. D. P. Servedio, G. Caldarelli
    http://xxx.lanl.gov/abs/cond-mat/0310503

- Board of Directors
  - M. E. J. Newman, S. H. Strogatz and D. J. Watts,
  - S. Battiston, E. Bonabeau and G. Weisbuch

Through this new description we can

- Discover new features
- Validate Models
**2C Stock Correlations**

\[
r_i(\tau) = \ln P_i(\tau) - \ln P_i(\tau - 1)
\]

Logarithmic return of stock \(i\)

\[
\rho_{i,j} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{\langle r_i^2 \rangle - \langle r_i \rangle^2} \sqrt{\langle r_j^2 \rangle - \langle r_j \rangle^2}}
\]

Correlation between returns (averaged on trading days)

\[
d_{i,j} = \sqrt{2(1 - \rho_{i,j})}
\]

Distance between stocks \(i, j\)

A tree (a graph with no cycle) can be constructed by imposing that the sum of the (N-1) distances is the minimum one.
Real Data from NYSE

Correlation based minimal spanning trees of real data from daily stock returns of 1071 stocks for the 12-year period 1987-1998 (3030 trading days). The node colour is based on Standard Industrial Classification system. The correspondence is:

- red for mining
- green for transportation, communications, electric, gas and sanitary services
- black for retail trade
- cyan for construction
- light blue for public administration
- purple for finance and insurance
- yellow for manufacturing
- magenta for wholesale trade
- orange for service industries

“Topology of correlation based…” http://xxx.lanl.gov/abs/cond-mat/0211546
G. Bonanno, G. C., F. Lillo, R. Mantegna
Troisieme Cycle Suisse Romande
Stat. Mech. of Networks-
Data from Capital Asset Pricing Model

In the model it is supposed that returns follow

\[ r_i(t) = \alpha_i + \beta_i r_M(t) + \varepsilon_i(t) \]

- \( r_i(t) \) = return of stock \( i \)
- \( r_M(t) \) = return of market (Standard & Poor’s)
- \( \alpha_i, \beta_i \) = real parameters
- \( \varepsilon_i \) = noise term with 0 mean

Correlation based minimal spanning trees of an artificial market composed by of 1071 stocks according to the one factor model.
The node colour is based on Standard Industrial Classification system. The correspondence is:

- red for mining
- green for transportation, communications, electric, gas and sanitary services
- black for retail trade
- cyan for construction
- light blue for public administration
- purple for finance and insurance
- yellow for manufacturing
- magenta for wholesale trade
- orange for service industries
Without going in much detail about degree distribution or clustering of the two graphs, we can conclude that:

the topology of MST for the real and an artificial market are greatly different.

Real market properties are not reproduced by simple random models.
•2C Portfolio Composition

Investors or Companies not traded at Borsa di Milano (Italy)
Companies traded at Borsa di Milano (Italy)
2C Portfolio Composition

Troisième Cycle Suisse Romande
Stat. Mech. of Networks
2C Portfolio Composition

CACB  Cascade Bank Corporate
FTIB  Fifth Third Bank Corporate
FFBC  First Financial Bank Corporation
NTRS  Northern Trust Corporation
SAFC  Safeco Corporation

Troisieme Cycle Suisse Romande
Stat. Mech. of Networks
2C Portfolio Composition
2C Portfolio Composition

In degree

Market Cap

Total Share

Weights

Troisieme Cycle Suisse Romande
Stat. Mech. of Networks
It is not only the topology that matters.

In this case as in many other graphs the weight of the link is crucial.

For every stock $i$ you compute this quantity.
The sum runs over the different holders

- If there is one dominating holder $SI$ tends to one
- If all the holders have a similar part $SI$ tends to $1/N$

For every guy $j$ you compute this quantity.
The sum at the denominator runs over the different holders of $i$
Then you sum on the different stocks in the portfolio
This gives a measure of the number of stocks controlled.
2C Stock correlation
FOOD CHAIN = sequence of *predation relations* among different living species sharing the same physical space (Elton, 1927):

- Flow of matter and energy from prey to predator, in more and more complex forms.
- The species ultimately feed on the abiotic environment (light, water, chemicals);
- At each predation, almost 10% of the resources are transferred from prey to predator.
A series of different interconnected food chains form a food web
**Trophic Species:**
Set of species sharing the same set of preys and the same set of predators (*food web* $\rightarrow$ *aggregated food web*).

**Trophic Level of a species:**
Minimum number of predations separating it from the environment.

- **Basal Species:** Species with no prey (B)
- **Top Species:** Species with no predators (T)
- **Intermediate Species:** Species with both prey and predators (I)

Prey/Predator Ratio = \( \frac{B+I}{I+T} \)
How to characterize the topology of Food Webs?

Graph Theory

Pamlico Estuary (North Carolina): 14 species

Aggregated Food Web of Little Rock Lake (Wisconsin)*: 182 species → 93 trophic species

* See Neo Martinez Group at http://userwww.sfsu.edu/~webhead/lrl.html
Unaggregated versions of real webs:

\[ P(k) \sim k^{-\gamma} \]


Aggregated versions of real webs:

Same qualitative behaviour of their unaggregated counterparts.

We look for other quantities!.
A spanning tree of a connected directed graph is any of its connected directed subtrees with the same number of vertices.

In general, the same graph can have more spanning trees with different topologies.
2D Food Webs Spanning Trees from data

St. Martin’s Island (Antilles):
44 species → 42 trophic species
224 links → 211 trophic links
(low taxonomic resolution)

Ythan Estuary (Scotland):
134 species → 123 trophic species
597 links → 576 trophic links
(taxonomic resolution: 88%)

Silwood Park (United Kingdom):
154 species → 83 trophic species
365 links → 215 trophic links
(taxonomic resolution: 100%)

Little Rock Lake (Wisconsin):
182 species → 93 trophic species
2494 links → 1046 trophic links
(taxonomic resolution: 31%)

Spanning Tree:
All edges directed from level $l_1$ to levels $l_2 \leq l_1$ are removed
Network of Interaction for the protein of Baker’s Yeast (*Saccharomyces Cerevisiae*)
How do growth and preferential attachment apply to protein networks?

- **Growth**: genes (that encode proteins) can be, sometimes, duplicated; mutations change some of the interactions with respect to the parent protein

- **Preferential attachment**: the probability that a protein acquires a new connection is related to the probability that one of its neighbors is duplicated; proportional to its connectivity

A. Vazquez et al., *ComPlexUs* 1, 38-44 (2003)
2D Two-hybrid method

The two hybrid method way of detecting protein interactions
2D More refined Models

With the solvation free energies taken from an exponential probability distribution $p(f) = e^{-f}$, we obtain

$$P(k) \sim k^{-2}$$

- The real network is random
- The detection method sees only pairs with large enough binding constants
- The binding constant is related to the solubilities of the two proteins
- Solubilities are given according to some distribution
2D Protein Interactions

Scale-Free Degree distribution

Scale-Free Betweenness $b(k)$

Troisieme Cycle Suisse Romande
Stat. Mech. of Networks-
• 2D Protein Interactions

Clustering per degree $c(k) \rightarrow$

$\leftarrow$ neighbors degree per degree $K_{nn}(k)$