Introduction to networks from a statistical point of view

Jean-Jacques Daudin, UMR AgroParisTech/INRA518

Statistics of Networks, Royal Statistical Society,
13 November 2007,
<table>
<thead>
<tr>
<th>Rank</th>
<th>Research Front</th>
<th>Core Papers</th>
<th>Citations</th>
<th>Citations / paper</th>
<th>Mean year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Complex networks</strong>, Statistical mechanics, Structure, Function, Evolution</td>
<td>3</td>
<td>3415</td>
<td>1138</td>
<td>2002.3</td>
</tr>
<tr>
<td>2</td>
<td>Current-driven magnetic domain wall motion</td>
<td>27</td>
<td>1385</td>
<td>51.3</td>
<td>2004.6</td>
</tr>
<tr>
<td>3</td>
<td>False discovery rate procedure</td>
<td>9</td>
<td>936</td>
<td>104</td>
<td>2003.3</td>
</tr>
<tr>
<td>4</td>
<td>Evolutionary game dynamics, <strong>Social networks</strong></td>
<td>12</td>
<td>382</td>
<td>31.83</td>
<td>2004.8</td>
</tr>
<tr>
<td>5</td>
<td>Porous plate using homotopy analysis method</td>
<td>12</td>
<td>359</td>
<td>29.92</td>
<td>2003.7</td>
</tr>
</tbody>
</table>
ISI WEB of Knowledge Essential Science Indicators Research Fronts rankings in Mathematics, sorted by Year

<table>
<thead>
<tr>
<th>Rank</th>
<th>Research Front</th>
<th>Core Papers</th>
<th>Citations</th>
<th>Citations / paper</th>
<th>Mean year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Exponential random graph (p) models, Curved exponential family models, Social networks</td>
<td>4</td>
<td>16</td>
<td>4</td>
<td>2006.7</td>
</tr>
<tr>
<td>2</td>
<td>Adaptive/flexible clinical trial design, Confirmatory clinical trials</td>
<td>5</td>
<td>17</td>
<td>5.6</td>
<td>2006</td>
</tr>
<tr>
<td>3</td>
<td>BOSE-EINSTEIN condensation</td>
<td>2</td>
<td>9</td>
<td>4.5</td>
<td>2006</td>
</tr>
<tr>
<td>4</td>
<td>CHEBYSHEV Series form, Computing real roots</td>
<td>3</td>
<td>15</td>
<td>5</td>
<td>2006</td>
</tr>
</tbody>
</table>
Attractive workshops in attractive places

INTERNATIONAL SCHOOL ON COMPLEXITY

5th Course

BIOLOGICAL NETWORKS

ERICE-SICILY: 9 - 11 OCTOBER 2006

Sponsored by the: • Italian Ministry of Education, University and Scientific Research • Sicilian Regional Government

TOPICS AND LECTURERS

Protein-protein interaction
Protein folding
Gene regulation
Metabolic pathways

• L. BARABÁSI, University of Notre Dame, IN, USA
• S. BORNHOLDT, Bremen University, D
• A. CAFLISCH, University of Zurich, CH
• A. EMILI, Toronto University, CDN
• T. FINK, Institute Marie Curie, Paris, F
• A. FLAMMINI, Indiana University, Bloomington, IN, USA
• P. LIÒ, Cambridge University, UK
• S. LEIBLER, Rockefeller University, New York, NY, USA
• L. LUISI, University of Rome 3, I
• D. MUELLER, ETH Zurich, CH
• E. RAVASZ, Los Alamos National Laboratory, NM, USA
• G. P. ROSSIINI, University of Modena, I
• D. SEGRÈ, Boston University, MA, USA
• E. SHAKHNOVICH, Harvard University, Boston, MA, USA
• G. STOLOVITZKY, IBM Research TJ Watson Center, USA
• P. UETZ, Karlsruhe University, D
• M. VENDORUSCOLO, Cambridge University, UK
• M. VERGASSOLA, Institute Pasteur, Paris, F
• M. VIDAL, Harvard University, Boston, MA, USA
• A. VESPIGNANI, Indiana University, Bloomington, IN, USA

More information about the «ETTORE MAJORANA» Foundation and Centre for Scientific Culture TO PAY A PERMANENT TRIBUTE TO GALILEO GALILEI, FOUNDER OF MODERN SCIENCE AND TO ENRICO FERMÎ, THE “ITALIAN NAVIGATOR”, FATHER OF THE WEAK FORCES ETTORE MAJORANA CENTENARY

EMFCSC PRESIDENT AND DIRECTOR OF THE CENTRE
Biological systems show emergent properties that are not readily explainable by the study of their constituent parts...

New mathematical formalisms, based on graph theory, are emerging in order to represent and study the underlying interaction networks present in the cell. The understanding of the evolution and organization of these networks is changing the way scientists look at biology. More specifically, the meeting will focus on:

- gene regulation
- metabolic pathways analysis
- protein-protein interactions
Complex networks = a speculative bubble or a true new scientific topic for statisticians?

1. The network representation of data and questions
2. Some basic definitions
3. Characterization of networks using topological motifs
4. Node clustering
5. Class prediction for the nodes
6. Statistical models for networks
7. Conclusions
Network representation of data and questions
An unusual data set structure

Usual i.i.d. structure

<table>
<thead>
<tr>
<th>item</th>
<th>$X_1$</th>
<th>...</th>
<th>...</th>
<th>$X_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$x_{11}$</td>
<td>...</td>
<td>...</td>
<td>$x_{1p}$</td>
</tr>
<tr>
<td>2</td>
<td>$x_{21}$</td>
<td>...</td>
<td>...</td>
<td>$x_{2p}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>$n$</td>
<td>$x_{n1}$</td>
<td>...</td>
<td>...</td>
<td>$x_{np}$</td>
</tr>
</tbody>
</table>

Structure for relational data

<table>
<thead>
<tr>
<th>item1</th>
<th>item2</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>$r_{12}$</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>$r_{13}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>$n-1$</td>
<td>$n$</td>
<td>$r_{n-1,n}$</td>
</tr>
</tbody>
</table>

- In the relational data set, the core information is the relation between two items.
- lines are not independent
- the data structure is similar to distance, similarity, covariance or correlation matrices
- the two types of structures may be combined: information about items + information about relations between them.
Representations of a correlation matrix by a graph

\[
\begin{pmatrix}
1 & 0.8 & 0.5 & 0.2 \\
1 & 0.1 & 0.1 \\
1 & 0.1 \\
1
\end{pmatrix}
\]

Correlation matrix

Valued graph

Binary graph using 0.15 threshold

Not readable if more than 10 items

Not readable if more than 10 items

Clusters and hubs appears, but the result depends on the threshold used and the representation software.

Other graphical representations are possible: colored intensity matrix, PCA...
Why do people represent data by a network (1)?

- Real networks do exist: electric, transport or www networks... They have been represented for a long time by virtual networks.
- Virtual network is a nice way for representing or even "modelling" many scientific phenomenons: social relations, metabolic pathways, chemical reactions...

FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).
Why do people represent data by a network (2) ?

- an overall representation of the interactions between many nodes
- the plot reveals the topology of the networks
- nodes may be colored, adding more information
Why do people represent data by a network (3)?

- nodes may be of different sizes, adding more information
- edges may be colored, adding more information
- a movie may report the evolution of the network in time.

2200 persons (nodes) from the Framingham Heart study. Circles with red borders = women, circles with blue borders = men.

The size of each circle is proportional to BMI. The color of the circles indicates the person's obesity status: yellow = obese person and green = nonobese person. The colors of the ties between the nodes indicate the relationship between them: purple denotes a friendship or marital tie and orange denotes a familial tie. N.A. Christakis et al. N Engl J Med, 2007, July
Choice of the level of detail of the information to be included in the (metabolic) network

Three possibilities:

- (a) include the catalyst Mg2+,
- (b) include the co-factors ADP, ATP,...
- (c) skeleton information.

Tools for visually exploring networks

- No unique representation
- Many tools and softwares
Questions induced by the network representation of the data: the focus is on topological questions (1)

Are there clusters composed of highly connected nodes? something like connected components in a weak sense.

- This question is related to the question of the existence of functional modules in biological networks, i.e. sub-networks which can produce outputs independently.

- the second rule of Descartes’s method is *Diviser chacune des difficultés ... en autant de parcelles qu’il se pourrait ... pour les mieux résoudre*. Each cluster may be studied independently and thus we are allowed to work on lower sized networks.

FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).
Questions induced by the network representation of the data: the focus is on topological questions(2)

Are there clusters composed of highly connected nodes? something like connected components in a weak sense.

- the rare edges between the clusters are the most important ones for the network’s connectivity.

- If a node A, with unknown properties, is in the same cluster that a well known node B, then we are tempted to attribute B’s properties to A.

FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).
Questions induced by the network representation of the data: the focus is on topological questions (3)

• How many hubs? They are the most important nodes in the network. If they are suppressed, the network’s connectivity is highly decreased.

• More generally, we are tempted to count some particular motifs (hub, triangle, square...). Some of them may be highly frequent so that we can characterize a network by its most frequent motifs.

• Resilience of the network’s connectivity to edge suppression.

FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).
Data and Questions for which the network representation seems less useful

- Weighted networks with more than 10 nodes
- Supervised classification of nodes (prediction of the function of proteins...)
- Supervised classification of edges.
Some basic definitions and properties about networks
Basic definitions

- A graph is $G = (V, E)$ where $V$ is a set of vertices (or nodes) and $E$ is a set of values corresponding to each couple of vertices $V$.
- If the values of $E$ are real, $G$ is valued and is called a *weighted* or *valued* graph.
- If the values of $E$ are in $(0, 1)$, $G$ is called a *direct graph*.
- If the values of $E$ are in $(0, 1)$ and $e_{ij} = e_{ji} \forall (i, j) \in V$, $G$ is called an *undirected graph*.
Basic definitions for undirected graphs

**Adjacency matrix** $X$ with $X_{ij} = 1$ if node $i$ is connected to node $j$ and else $X_{ij} = 0$.

**Degree of node** $i$ $K_i = \sum_j X_{ij}$

**Distance between** $i$ **and** $j$ length of the shortest path between $i$ and $j$

**clique** a clique is a set of vertices, such that for every two vertices, there exists an edge connecting them.

![Graph Image]

$$X = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$K_1 = 2, K_2 = 3, K_3 = 2, K_4 = 2, K_5 = 1$

$D(3,5) = 3$

$(1,2,3)$ is a clique
Characterization of a network by the counts of some topological motifs
## Topological motifs

<table>
<thead>
<tr>
<th>$z(x)$</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$-degree</td>
<td>number of nodes with degree $k$</td>
</tr>
<tr>
<td>Triangle</td>
<td>number of triangles</td>
</tr>
<tr>
<td>$V$</td>
<td>number of $V$</td>
</tr>
<tr>
<td>$C$</td>
<td>$3 \cdot \text{(number of triangles)} / \text{number of } V$</td>
</tr>
<tr>
<td>Squares</td>
<td>number of squares</td>
</tr>
<tr>
<td>$k$-star</td>
<td>number of nodes with $k$ edges with unconnected endpoints</td>
</tr>
<tr>
<td>geodesic</td>
<td>length of the longest distance between two nodes</td>
</tr>
<tr>
<td>edge count</td>
<td>number of edges</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Empirical distribution of the degrees of a chemical reaction network

V. Lacroix, M-F Sagot
Degrees pdf "scale free", popular in the physicist community 5 years ago

\[ P(K = k) = c(\rho) k^{-(\rho)} \]

with \( k \in \mathbb{N}, k > k_0, \rho > 1 \) et

\[
c(\rho) = \left[ \sum_{k > k_0} k^{-(\rho)} \right].
\]

Scale free \( \iff \frac{P(K=\alpha k)}{P(K=k)} = \frac{(\alpha k)^{-(\rho)}}{k^{-(\rho)}} = \alpha^{-\rho} \) do not depend on \( k \).

\[
\log(P(K = k)) = \log(c(\rho)) - \rho \log k
\]

log-log plot \( y-axis: \log(\hat{P}(K = k)), x-axis: \log k. \)
Log-Log-Plot for the reaction network

Graphe d’interaction des réactions dans E. Coli
Scale-free, an ubiquitous pdf? Newman 2002

Introduction

Figure 1: Measured degree distributions for a number of different networks. (a) Physical connections between autonomous systems on the Internet, circa 1997 (Faloutsos et al., 1999).
(b) A 200 million page subset of the World-Wide Web, circa 1999 (Broder et al., 2000).
The figure shows the out-degree of pages, i.e., numbers of links pointing from those pages to other pages.
(c) Collaborations between biomedical scientists and between mathematicians (Newman, 2001b,d).
(d) Collaborations of film actors (Amara et al., 2000).
(e) Co-occurrence of words in the English language (i Cancho and Solé, 2001).
(f) Board membership of directors of Fortune 1000 companies for year 1999 (Newman et al., 2001).
Building the network by preferential attachment lead to scale-free networks

Barabasi 2002

Two basic mechanisms

• Growth
• Preferential attachment to nodes with high degree.
• Other pdf fit the empirical distribution of the degrees better than the scale-free pdf,
• The degree distribution is a poor information on the network topology.
• Samples of scale-free networks are not scale-free.
Triangle Coefficient (C)

\[ C = P(T/V) \]

T is a triangle and V is a "V".

The degree distribution and the C are partial and quite poor statistics for characterizing the network.
Characterization of a network by its degrees and C. Barabasi et al.
Recent developments of Exponential Random Graph Models, ERG or $p^*$ models, from Robins, Snijders, Handcock, Pattison 2007

$$P(X = x) = \frac{e^{\theta' z(x)}}{C(\theta)}$$

- $\theta$ is a $p$-vector of parameters
- $z(x)$ is a $p$-vector of motifs statistics.

A network is characterized by $\hat{\theta}$. This allows comparison between networks, by comparing their parameters estimates $\hat{\theta}$. Saul et al. Bioinformatics 2007, compared 43 metabolic networks using ERG models.
Exceptional motifs in a network

Fig. 2. Two sets of instances of the stochastic bi-fan motif in the *S. cerevisiae* regulatory network (7). (A) Instances related to the high-osmolarity glycerol (HOG) response pathway. (B) Instances related to the nitrogen regulation process. (C) The “backup” mechanism in the HOG pathway. (D) The “balancing” mechanism in the nitrogen regulation process.
Exceptionally frequent motifs in a network

209 motifs "bi-fan" in a directed gene regulation network of *E. Coli*
Is the count exceptional? Shen-Orr 2002

The results depend on the reference model, see Picard 2007.
Nodes clustering
Clustering of nodes

Santa Fe Institute collaboration network: the nodes of the network represent scientists from the Santa Fe Institute and an edge is drawn between two nodes if the corresponding scientists have coauthored at least one publication during the calendar year 1999 or 2000.
Clustering methods

Hierarchical clustering: a well-known set of algorithms for a similarity matrix, many criteria proposed.

Spectral clustering: k-means on the space associated to the \( q \)-lowest eigenvalues of the Laplacian of the network, \( D-X, \) with \( D = diag(sum(X)) \) tutorial from U. Von Luxburg, Stat Comput, 2007.

Markov Clustering Algorithm: MCL simulates a flow on the graph by calculating successive powers of \( X. \) At each iteration an inflation step is applied to enhance the contrast between regions of strong or weak flow in the graph. The process converges towards a partition of the graph, with a set of high flow regions separated by boundaries with no flow, package from Von Dongen.

Edge-betweenness clustering: Divisive method using the edge betweenness EB, (for an edge): the number of shortest paths that pass through the edge. The edge with the highest EB is removed, Girvan & Newman.
Statistical models for clustering

Discrete latent variable, Z

\[ P(X_{ij} = 1 / Z_i = q, Z_j = l) = \pi_{ql}, \]

packages BLOCKS 1.6 (Snijders et al.) and ERMG 05, (Robin et al.), the clusters obtained are not necessarily highly connected within clusters and poorly connected between clusters.

Continuous latent variables, Z

\[
\frac{P(X_{ij} = 1 / Z_i = z_i, Z_j = z_j)}{P(X_{ij} = 0 / Z_i = z_i, Z_j = z_j)} = e^{\beta(z_i - z_j)}
\]

where Z has a multivariate gaussian mixture pdf, package latentnet, Handcock et al.

Estimation methods: MCMC or variational ML.
Clusters in the model using a discrete latent variable

<table>
<thead>
<tr>
<th>Description</th>
<th>Graph</th>
<th>$Q$</th>
<th>$\Pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erdos</td>
<td><img src="image1.png" alt="Graph" /></td>
<td>1</td>
<td>$p$</td>
</tr>
</tbody>
</table>
| Star                         | ![Graph](image2.png) | 4   | \[
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0
\]
| cluster in usual sense       | ![Graph](image3.png) | 2   | \[
1 & \varepsilon \\
\varepsilon & 1
\]
Continuous latent variables and Principal Coordinate Strategy

<table>
<thead>
<tr>
<th>Relational data</th>
<th>PCO</th>
<th>Usual data structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>item1</td>
<td>item2</td>
<td>R</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>r_{12}</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>r_{13}</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>n − 1</td>
<td>n</td>
<td>r_{n−1,n}</td>
</tr>
</tbody>
</table>

Then, it is easy to merge the individual data set X and the latent data set Z because they have the same structure. This strategy, combined with kernel methods, has been used by Vert and Yamanishi for system biology.
Class prediction for nodes
Prediction of the function of a protein

Figure 1  Extent of annotation of proteins in model species. For each species, the charts give the fractions and numbers of annotated and unannotated proteins, according to the three ontologies of the GO annotation. The numbers are based on the Entrez Gene and the WormBase databases as of September 2006.
Function prediction of a protein: basic idea

Figure 3  Correlation between protein functional distance and network distance. X-axis: distance in the network. Y-axis: average functional similarity of protein pairs that lie at the specified distance. The functional similarity of two proteins is measured using the semantic similarity of their GO categories (Lord et al, 2003).
Direct prediction versus prediction after clustering

Figure 2  Direct versus module-assisted approaches for functional annotation. The scheme shows a network in which the functions of some proteins are known (top), where each function is indicated by a different color. Unannotated proteins are in white. In the direct methods (left), these proteins are assigned a color that is unusually prevalent among their neighbors. The direction of the edges indicates the influence of the annotated proteins on the unannotated ones. In the module-assisted methods (right), modules are first identified based on their density. Then, within each module, unannotated proteins are assigned a function that is unusually prevalent in the module. In both methods, proteins may be assigned with several functions.
Direct prediction using Markov Random Fields

\[ P[L(i) = 1/L(V(i))] = f[\log(\pi/(1-\pi)) + \beta N(i, 1) + \alpha [N(i, 1) - N(i, 0)] - N(i, 0)] \]

with

- \( L(i) \) is the label of node \( i \),
- \( N(i, 1) \) (resp. \( N(i,0) \)) is the number of neighbors of \( i \) with label 1 (resp. 0),
- \( \pi \) is the marginal probability of labels 1
- \( V(i) \) is the set of neighbors of \( i \),
- \( f(t) = 1/(1 + exp(-t)) \)
Statistical models for networks
Basic model for random networks, "Erdös-Rényi"

\[ \forall (i, j), P(X_{ij} = 1) = p. \]

\( X_{i,j} \) are iid Bernoulli\((p)\).

Historically and theoretically very important model, but too simple in practice.

Does not fit real data in any domain.
More probabilistic models for networks are necessary to:

1. summarize the information contained in one network and compare networks
2. have a reference model to say if a particular motif is exceptional
3. predict the label of a node
4. predict the value of an edge
5. create artificial networks by simulation
Complex networks = a speculative bubble or a true new scientific topic for statisticians?

• High level of "buzz" has interested many researchers on this topic → new models and new results

• In many domains (system biology, social sciences...) new data sets contain both individual and relational informations.

Example of the prediction of the function of proteins:

1. PPI network (relational data)
2. Blast score of AA or cDNA sequence (individual and relational data)
3. Gene Ontology (individual data)
4. Microarray expression data (individual data)

• The network plot is sometimes nice and fruitful and sometimes unnecessary and misleading.
Future challenges

- Weighted networks \((r_{ij} \in \mathbb{R})\)
- Multivariate relational data

Informations on items

<table>
<thead>
<tr>
<th>item</th>
<th>(X_1)</th>
<th>...</th>
<th>...</th>
<th>(X_p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(x_{11})</td>
<td>...</td>
<td>...</td>
<td>(x_{1p})</td>
</tr>
<tr>
<td>2</td>
<td>(x_{21})</td>
<td>...</td>
<td>...</td>
<td>(x_{2p})</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>(n)</td>
<td>(x_{n1})</td>
<td>...</td>
<td>...</td>
<td>(x_{np})</td>
</tr>
</tbody>
</table>

Informations on relations between items

<table>
<thead>
<tr>
<th>item1</th>
<th>item2</th>
<th>(R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>(r^{(1)}<em>{12}, r^{(2)}</em>{12}, \ldots, r^{(q)}_{12})</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>(r^{(1)}<em>{13}, r^{(2)}</em>{13}, \ldots, r^{(q)}_{13})</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>(n-1)</td>
<td>(n)</td>
<td>(r^{(1)}<em>{n-1,n}, \ldots, r^{(q)}</em>{n-1,n})</td>
</tr>
</tbody>
</table>

- Individual and relational data indexed by the time \(t\).
References in the domain of System Biology


• Sharan R., Ulitsky I. and Shamir R. *Network-based prediction of protein function*, Molecular Systems Biology, 3, Article 88, 2007


References in the physicist community

- A. Barabasi, Depart. of Physics, University of Notre Dame, http://www.nd.edu/alb/
- P.E. Newman, Depart. of Physics, University of Michigan, http://www-personal.umich.edu/mejn/
Exceptional motifs counts in networks

- *An extended transcriptional regulatory network of E. Coli and analysis of its hierarchical structure and network motifs*, NAR, 2004,32,22, 6643-6649
- R. Jiang, Z. Tu, T. Chen and F. Sun *Network motif identification in stochastic networks* PNAS, juin 2006
Models for graphs

• Barabasi, http://www.nd.edu/~alb/


Clustering of nodes of networks


Thank you for your attention