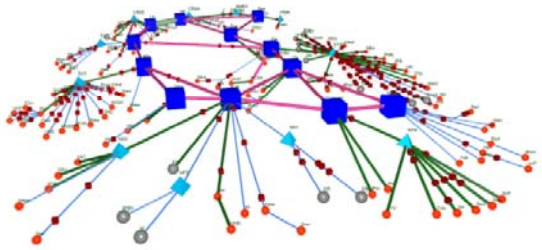
Friendship Network Formation

Albert Diaz-Guilera
http://albert.diaz.guilera.googlepages.com

Outlook

- Networks
  - Generalities
  - Complex networks in nature and society
  - Characterizing networks
  - Simple models
- Friendship Social Networks
  - General characteristics
  - Social networks data
  - Models

Networks



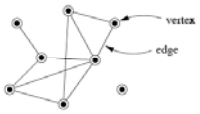
http://au.youtube.com/watch?v=zK1Cb9qj3qQ

Why is network anatomy important

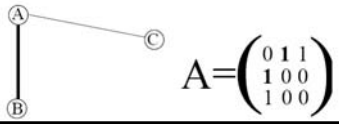
- Structure always affects function
- The topology of social networks affects the spread of information
- Internet
  - + access to the information
  - - electronic viruses

Representations

- Network (graph): nodes (vertices) + links (edges)



- Network (matrix): adjancency matrix

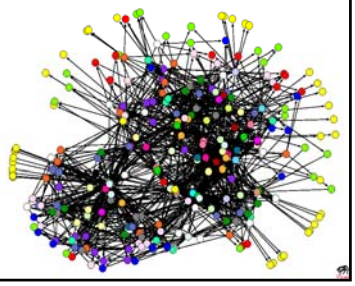


How are the networks we find?

- Nodes and links: binary
- Bipartite: two types of nodes
- Directed: nonsymmetric adjacency matrices
- Weighted: the links can have strength

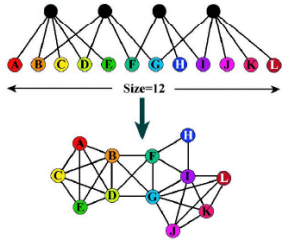
Common network

- Neural network of C. Elegans (a worm)
- Nodes = neurons
- Links = synapses



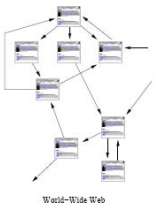
Bipartite graph

- Coauthorship network



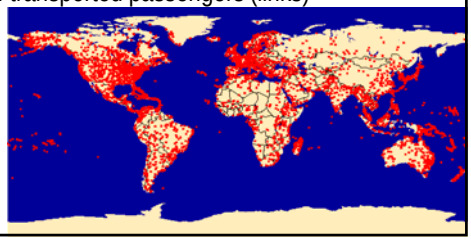
Directed relationships

- Sometimes relational data has a direction
- The adjacency matrix is not symmetric
- Examples:
  - links to web pages
  - information
  - cash flow
  - trust networks



Weighted networks

- Nodes and/or links can have different strengths
- Population of cities (nodes)
- Number of transported passengers (links)

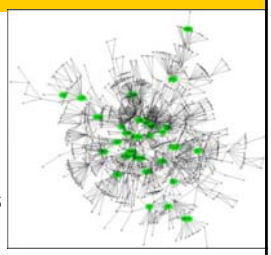


Complex networks in nature and society

- NOT regular lattices
- NOT random graphs "simple" mathematical analysis
- Huge databases and computer power

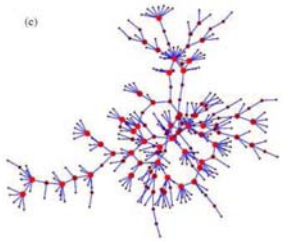
Social I: Networks of collaboration

- Through collaboration acts
- Examples:
  - movie actor
  - board of directors
  - musicians
  - scientific collaboration networks (MEDLINE, Mathematical, neuroscience, e-archives,..)



Social networks (II)

- Email
- Trust networks
- Sexual contacts

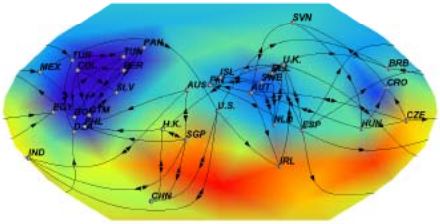


Economical systems

- Relations between stocks
- Relations between economic agents
- World Trade Web

Directly measurable connectivities

- World Trade Web (manufactured metal)

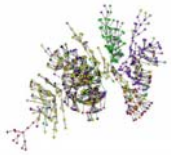


Indirect measures

- Correlations between dynamical properties: returns, GDP,...

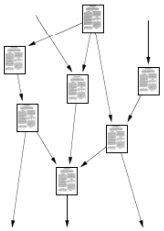
$$\rho_{ij}(\Delta t) = \frac{\langle p_i p_j \rangle - \langle p_i \rangle \langle p_j \rangle}{\sqrt{(\langle p_i^2 \rangle - \langle p_i \rangle^2)(\langle p_j^2 \rangle - \langle p_j \rangle^2)}}.$$

$$d_{ij} = \sqrt{2(1-\rho_{ij})}$$

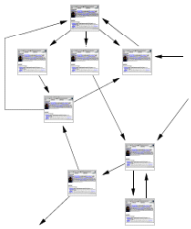


Information networks

- The World-Wide Web
- Citation networks



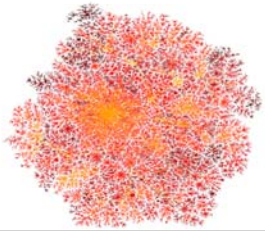
citation network

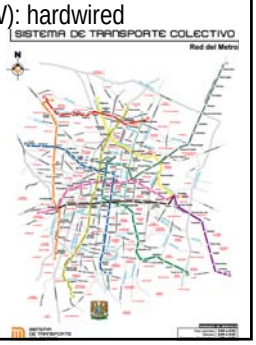


World–Wide Web

Technological Networks

- Internet (not the same than WWW): hardwired
- Power grid
- Transportation networks





Characterizing networks: Topological properties

- Degree distribution
- Clustering
- Shortest paths
- Betweenness
- Correlations

Degree (microscopic scale)

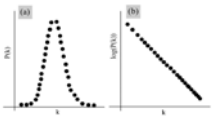
- Number of links that a node has
- It corresponds to the local centrality in social network analysis
- It measures how important is a node with respect to its nearest neighbors

Degree distribution (macroscopic scale)

- Gives an idea of the spread in the number of links the nodes have
- P(k) is the probability that a randomly selected node has k links

What should we expect?

- In regular lattices all nodes are identical
- In random networks the majority of nodes have approximately the same degree



- Real-world networks: this distribution has a power tail

$$P(k) \approx k^{-\gamma}$$ "scale-free" networks

Clustering

- Cycles in social network analysis language
- Circles of friends in which every member knows each other

Clustering coefficient

- Clustering coefficient of a node

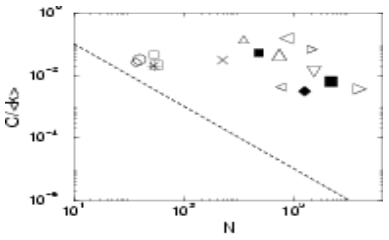
$$C_i = \frac{E_i}{k_i(k_i-1)/2}$$

- Clustering coefficient of the network

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

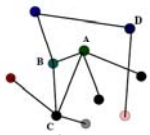
What happens in real networks?

- The clustering coefficient is much larger than it is in an equivalent random network



Distance between two nodes

- Number of links that make up the shortest-path between two nodes

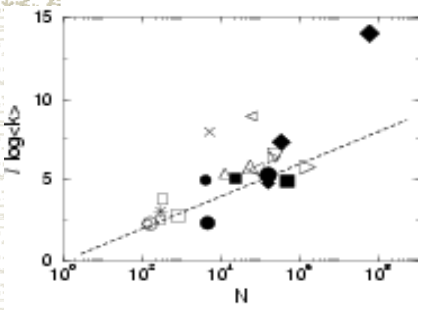


- Centrality: nodes that are “close” to many other nodes in the network.
- Global centrality: defined as the sum of minimum distances to any other nodes in the networks

Global centrality of the whole network?

Mean shortest path = average over all pairs of nodes in the network

What do we find?



Betweenness

- Measures the "intermediary" role in the network
- It is a set of matrices, one for ach node

B_{ij}^k Ratio of shortest paths bewteen i and j that go through k

$0 \le B_{ij}^k \le 1$ There can be more than one geodesic between i and j

- $B_k = \sum_{ij} B_{ij}^k$

is a measure of the centrality, in terms of flow, of node k

Correlations

- Degree correlations: expected degree of the neigbors of a node as a function of its degree

$$k_{\mathrm{nn}}(k) = \sum_{k'} k' P(k'|k)$$

Models of complex networks

- The Erdös-Rényi (ER) random graph model
- The Watts-Strogatz (WS) "small-world" model
- The Barabasi-Albert (BA) "scale-free" model

Erdös-Renyi random model

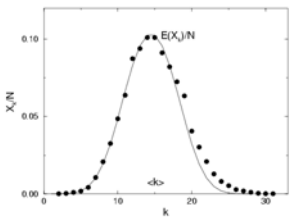
- start with N nodes, every pair of nodes being connected with probability p
- The total number of links, n, is a random variable
  - E(n)=pN(N-1)/2
- The degree of a node follows a binomial distribution

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

Large N

- Poisson distribution:

$$P(k_i = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$



Mean shortest path

$$l_{\text{rand}} \approx \frac{\ln N}{\ln < k >} \approx \frac{\ln N}{\ln pN}$$

Clustering coefficient

- Probability that two nodes are connected (given that they are connected to a third)?

$$C_{rand} = p = \frac{< k >}{N}$$

$$\frac{C_{rand}}{< k >} \approx \frac{1}{N}$$ while it is constant for real networks

Watts- Strogatz small-world model

- Small world: the average shortest path length in a real network is small
- Six degrees of separation (Milgram, 1967)
- Local neighborhood + long-range friends
- A random graph is a small world

Networks in nature (empirical observations)

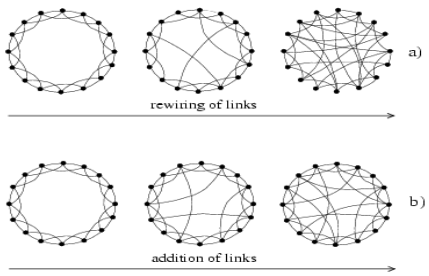
$$l_{\text{network}} \approx \ln(N)$$

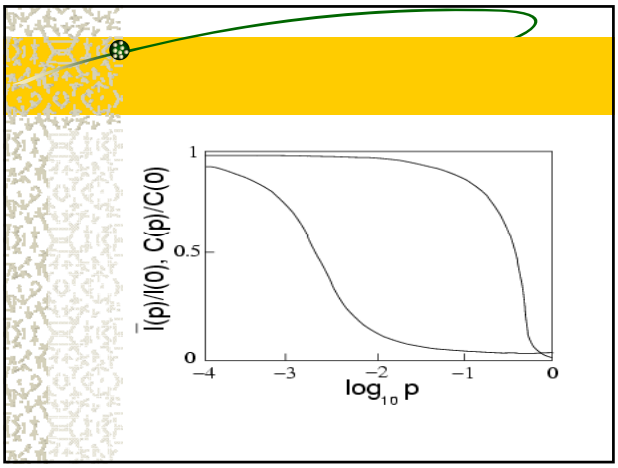
$$C_{\text{network}} >> C_{\text{random graph}}$$

Model proposed

- Crossover from regular lattices to random graphs
- Tunable
- Small world network with (simultaneously):
  - Small average shortest path
  - Large clustering coefficient (not obeyed by ER-RG)

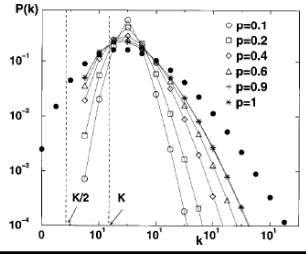
Two ways of constructing





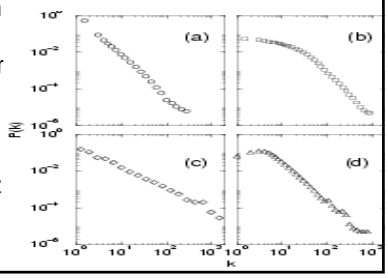
Degree distribution

- p=0 delta-function
- p>0 broadens the distribution
- Exponential decay



Barabasi-Albert: scale-free network

- Many large networks are scale free
- The degree distribution has a power-law behavior for large k (far from a Poisson distribution)
- Random graph theory and the Watts-Strogatz model cannot reproduce this feature



Barabasi-Albert model (1999)

- Two generic mechanisms common in many real networks
  - Growth (www, research literature, ...)
  - Preferential attachment (idem): attractiveness of popularity
- The two are necessary

Growth

- t=0, m_0 nodes
- Each time step we add a new node with m ($\leq m_0$) edges that link the new node to m different nodes already present in the system

Preferential attachment: rich gets richer

- When choosing the nodes to which the new connects, the probability Π that a new node will be connected to node *i* depends on the degree k_i of node *i*

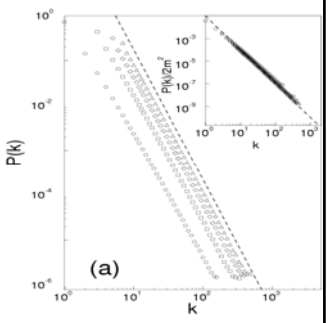
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

Linear attachment (more general models)
Sum over all existing nodes

Numerical simulations

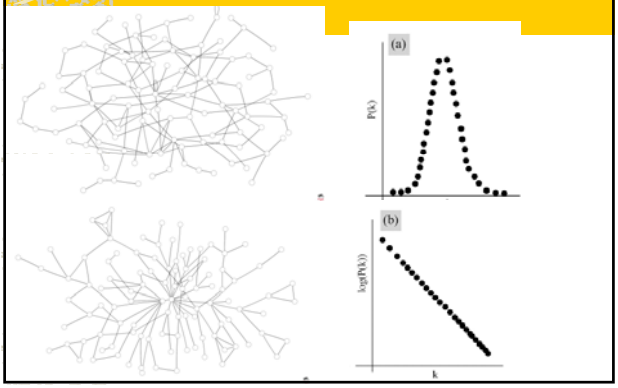
- Power-law $P(k) \approx k^{-\gamma}$ $\gamma_{SF}=3$
- The exponent does not depend on m (the only parameter of the model)
- Calculation: continuity equation

$$P_k = \frac{2m(m+1)}{k(k+1)(k+2)}$$



γ=3. different m's. P(k) changes. γ not

Hubs



The mesoscale: communities

- Existence of communities or modules in networks
- Technical issue: finding the best partition
- Management issue: finding meaningful partitions

Communities

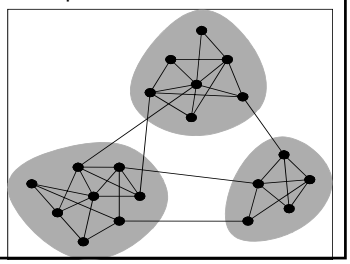
- Social networks are formed by communities
- Functional modules in biological networks
- Many different origins: political, religious, function, …

Technical issue

- We have to identify the communities
- How many possible partitions into communities?
- NP problem to find the best one

Communities: intuitive picture

- Definition: subsets of nodes that are more densely linked, when compared with the rest of the network



Partition

- A partition is a division of the network into groups, communities or clusters
- The question is: Which of all possible partitions is the best?
- NP problem
- Community detection:
  - From computer scientists
  - To statistical physicists (Girvan-Newman, PNAS 99, 7821, 2002)

Quantifying a partition

•Modularity:

$$Q = \sum_i (e_{ii} - a_i^2)$$

•e_{ij}: fraction of total links starting at a node in partition *i* and ending at a node in partition *j*

•a_i: fraction of links connected to *i*

•a_i^2: number of intracommunity links

Identifying communities

- Identifying what communities are
- How a group is created, organized, grown, …

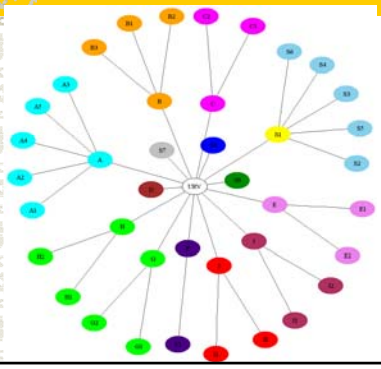
Data

- Social networks
- Friendship networks

Social networks

- Email at URV
- Pretty – Good – Privacy (PGP)

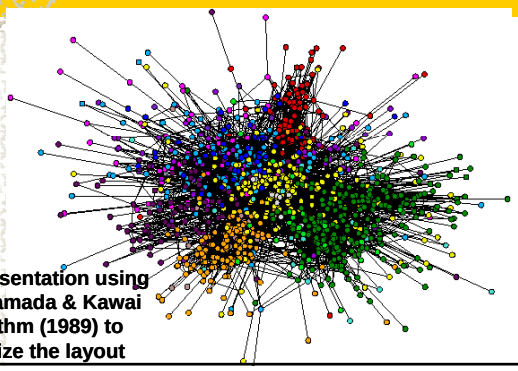
URV



Data acquisition to construct the e-mail network of the URV

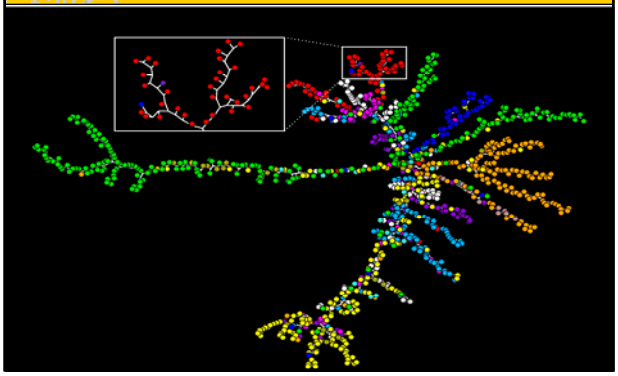
- Node => e-mail address
- Link => bidirectional e-mails between nodes (undirected graph)
- Number of users approx. 1700 (professors, technicians, administrators, graduate students)
- We consider only e-mails sent within the University during the first 3 months of 2002 (stable network)
- Non “spam” mail: (neglect >50 recipients)

Email at URV



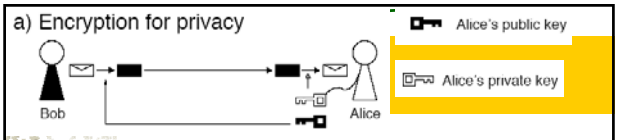
Representation using the Kamada & Kawai algorithm (1989) to optimize the layout

Communities in URV



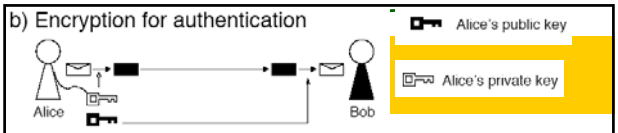
PGP web of trust

- Pretty-Good-Privacy (PGP) encryption system
- Ensures privacy and authentication



A user can encrypt a message so that only the desired addressee can read it. For example, Bob wants to send a message to Alice and make sure that only she will be able to read it. He then takes Alice's public (black) key and encrypts the message, which will travel encrypted through the Internet. Once Alice receives the message, she can decipher it using her private (white) key. Although a third person may be able to intercept the message, he would not be able to decipher it because he would need Alice's private key.

b) Encryption for authentication



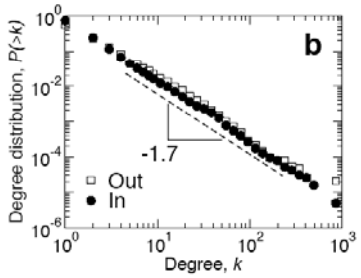
A user can digitally authenticate a message so that the recipient of the information can verify the authenticity of the information's origin. For example, Alice wants to send a message to Bob and wants to make sure that he knows with absolute certainty that she is the sender of the message. To this end, she encrypts the message with her private key. If Bob is able to decipher it using Alice's public key, then the message must have originated with Alice's private key and, moreover, the message must be intact.

The web of PGP signatures

- Use keys of users you know directly, over the phone, for instance
- A central authority that certifies signatures?
- PGP web of trust users sign public keys certifying a users is who it claims to be
- This creates a trust relation
- A directed network of trust

The directed network: connectivity distribution

- July 2001:



191,548 keys
286,290 signs.

$P(> k) \propto k^{-\alpha}$

$\alpha_{in} = -1.8 \pm 0.2$

$\alpha_{out} = -1.7 \pm 0.2$

Friendship networks data

- Friendship of graduate students in Japan
- Instant messaging
- Erasmus and Eurovision
- Students in US

Japanese graduate students

Table 1. Observed data

Organization	School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST).
Participants	Graduate students who have just entered the shool.
Period	Every week from April 2005 to May 2005. 8 weeks for group A and 7 weeks for group B.
Group	People who are attending a lecture form a group. Two groups (two lectures) are investigated. The number of members in group A and group B are 26 and 16, respectively. The groups do not overlap.
Method	Questionnaire that enquires about the friendship among the people in a particular group.

Figure 1. Start-up network (left) and the network at the end of the survey (right) in group A.

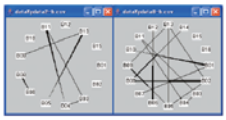
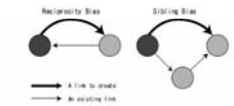


Figure 2. Start-up network (left) and the network at the end of the survey (right) in group B.

Biases

- Reciprocity and sibling: links created based on the network structure. Important for the model



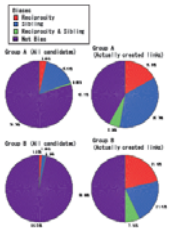
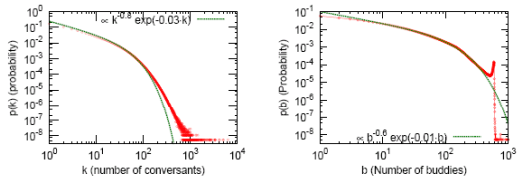


Figure 8. Ratio of types of biased links

Instant messaging: a world-wide networks

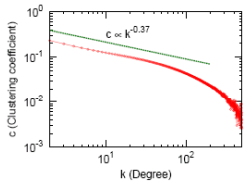
- **Presence events:** These include login, logout, first ever login, add, remove and block a buddy, add unregistered buddy (invite new user), change of status (busy, away, be-right-back, idle, etc.). Events are user and time stamped.
- **Communication:** For each user participating in the session, the log contains the following tuple: session id, user id, time joined the session, time left the session, number of messages sent, number of messages received.
- **User data:** For each user, the following self-reported information is stored: age, gender, location (country, ZIP), language, and IP address. We use the IP address to decode the geographical coordinates, which we then use to position users on the globe and to calculate distances.

Statistics of the communication network

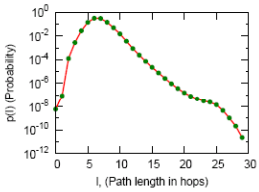


(a) Communication (b) Buddies

Figure 18: (a) Degree distribution of communication network (number of people with whom a person communicates). (b) Degree distribution of the buddy network (length of the contact list).



(a) Clustering



(a) Diameter

Sexual interaction networks

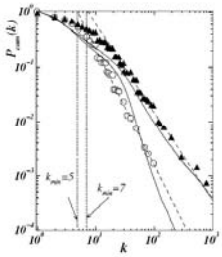


Fig. 4. Cumulative distribution of sexual partners in a network of heterosexual contacts extracted from reference [6], where male (triangles) and females (circles) distributions are plotted separately, with a total of 2810 persons. Solid lines indicate the simulations when plotting the distributions at the same stage $n = 0.2N$, starting with a population composed by 58% of females and 42% of males. Here $N = 10^6$, $s = 0$.

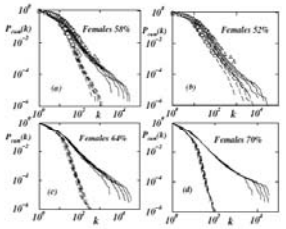


Fig. 5. Distributions of sexual partners in a network of heterosexual contacts using different amounts of females: (a) 58% as in Figure 4, (b) 52%, (c) 64% and (d) 70%. In each plot solid and dashed lines indicate the cumulative distributions of males and females respectively, for five different realizations. Clearly, the exponent of the power-law tail of the distributions decreases when the percentage of females or males increases (see dotted lines). Same conditions as in Figure 4 were used.

Friendship in American schools

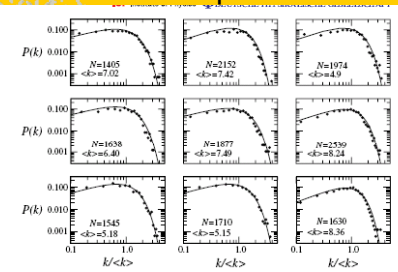
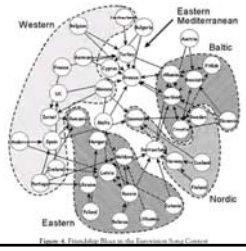


Figure 3. Degree distributions of nine different schools (symbols) from an in-school questionnaire involving a total of 90 118 students which responded in a survey between 1994 and 1995. Each school comprises a number N of interviewed students and from their questionnaires an average number $\langle k \rangle$ of acquaintances is extracted. With solid lines, we represent the fit obtained with a Brody distribution, equation (2), whose parameter value is computed in figure 4.

Other data

- Friendship in Erasmus students
- Eurovision contest as a friendship network: who votes who?



Models

- Model for Japanese graduate schools
- Model based on collision
- Model based on social distance

Japanese model: biases

1. Add n_0 nodes to the network.
2. Add m links to the network based on the probability calculated by equation 6.
3. Add n_t nodes having no links to the network if necessarily.
4. Repeat steps 2 to 3 an arbitrary number of times.

$$\begin{aligned} cad(x \to y) &= \theta_{rep} * rep(x \to y) + \\ &\quad \theta_{stb} * sib(x \to y) + \\ &\quad \theta_{nob} * nobias(x \to y), \\ &\quad \theta_{rep} \geq 0, \theta_{stb} \geq 0, \\ &\quad \theta_{nob} \geq 0, \\ &\quad \theta_{rep} + \theta_{stb} + \theta_{nob} = 1 \quad (5) \\ P(x \to y) &= cad(x \to y) / \\ &\quad \sum_{i \in X} \sum_{j \in Y} cad(i \to j) \quad (6) \end{aligned}$$

where $rep(x \to y)$, $sib(x \to y)$, and $nobias(x \to y)$ denote the reciprocity bias, sibling bias, and absence of bias from nodes x to y, respectively. θ_{rep}, θ_{stb}, θ_{nob} are the simulation parameters.

Model for mobile agents

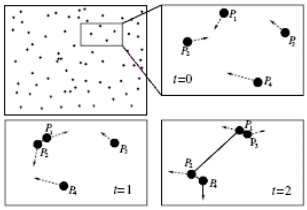


Figure 1. Illustration of the two-dimensional mobile agents system. Initially there are no connections between nodes and nodes move with some initial velocity v_0 in a randomly chosen direction (arrows). At $t = 1$ two nodes, P_1 and P_2 collide and a connection between them is introduced (solid line), velocities are updated increasing their magnitude and choosing a new random direction. At $t = 2$ two other collisions occur, between nodes P_2 and P_4 and between nodes P_1 and P_3. In this way a network of nodes and connections between them emerges as a straightforward consequence of their motion (see text).

Preferential

Assuming that large numbers of acquaintances tend to favour the occurrence of new contacts, the velocity should increase with degree k, namely

$$\vec{v}(k_i) = (\bar{v}k_i^{\alpha} + v_0)\vec{\omega}, \quad (1)$$

- Correctly parameterized it reproduces results from sexual interactions, American schools, ...

A model based on social distance

- Relations are established between individuals that feel “close”
- Assign a position to each individual = a set of values in a number of social dimensions

1 social dimension

- Position of individual i in social space h_i
- Define a density $\rho(h)$
- Probability of establishing a relation decreases with social distance

$$r_n(h_i^n, h_j^n) = \frac{1}{1 + \left[b_n^{-1} d_n(h_i^n, h_j^n)\right]^{\alpha_n}}$$

- α measures the degree of homophily

Some results

- Homogeneous density $\rho(h)=1/h_{max}$
- Average degree

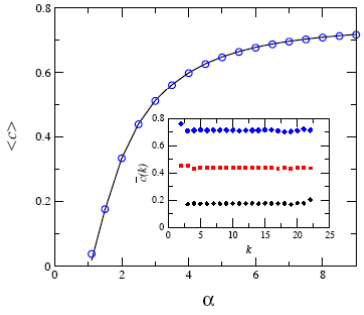
$$\langle k \rangle = \frac{2\delta b \pi}{\alpha \sin \pi/\alpha} \qquad \delta = N/h_{max}$$

- Clustering coefficient

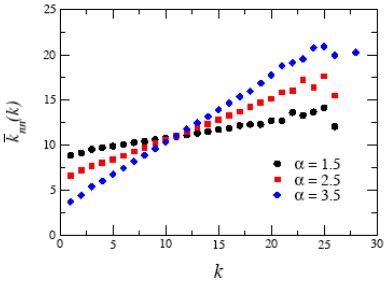
$$\langle c \rangle = \frac{\alpha^2}{4\pi^2} f(\alpha) \sin^2 \frac{\pi}{\alpha}$$

$$f(\alpha) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{dxdy}{(1+|x|^{\alpha})(1+|x-y|^{\alpha})(1+|y|^{\alpha})}$$

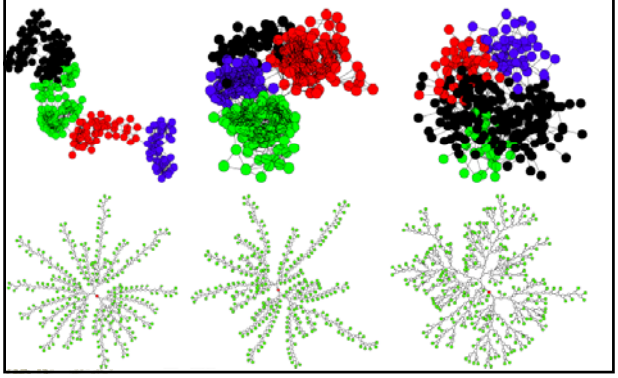
Clustering coefficient



Assortativity: degree correlations



Communities



Bibliography

- A.-L. Barabasi & R. Albert: Statistical Mechanics of Complex networks, Rev.Mod.Phys. 74 (2002) 47.
- A. Diaz-Guilera:*Complex networks: Statics and Dynamics*, AIP Conference Proceedings -- February 9, 2007 -- Volume 885, pp. 107-128.
- A. Arenas, L. Danon, A. Díaz-Guilera, P.M. Gleiser and R. Guimerà, Community analysis in complex networks, European Physics Journal B 38 (2004) 373
- J. Leskovec & E. Horvitz, Planetary-Scale Views on an Instant-Messaging Network, arxiv: arXiv:0803.0939v1

Bibliography (II)

- A. Arenas, M. Boguña, A. D-G., R. Pastor-Satorras, *Models of social networks based on social distance attachments*, Physical Review E 70 (2004) 056122.
- T.Nakada, Y.Kato, S. Kunifuyi, A study on the dynamics of friendship network formation using a directed network model, preprint.

Bibliography (III)

- A. de Federico de la Rúa, A., L'analyss longitudinal de reseaux sociaux totaux avec SIENA- Methode, discussion et application. *Bulletin de Méthodologie Sociologique. Nº 84, octobre, pp. 5-39.*
- A. Dekker, The Eurovision Song Contest as a « Friendship» Network, *Connections 28 (2008) 59*
- P.J.Lind and H.J. Herrmann, *New approaches to model and study social networks, New.J.Phys. 9 (2007) 228.*