Introduction to complex networks

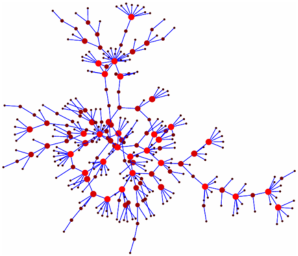
Flavia Bonomo

October 20, 2009

What is a network?

Network: a collection of entities that are interconnected with links. For example:

- people that are friends
- computers that are interconnected
- web pages that point to each other
- proteins that interact



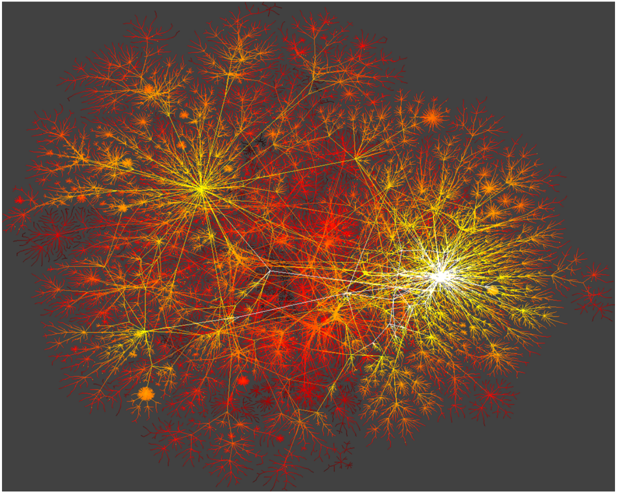
In terms of **graph theory**, the entities are called vertices and the links edges.

What is a complex network?

Large graphs of real life are called complex networks. Some of the main questions about them are the following:

- What are the statistics of real life networks?
- Can we explain how the networks were generated?

Example: the Internet graph



More examples

- Social networks:
  - networks of acquaintances
  - collaboration networks
  - phone-call networks
- Technological networks:
  - the Internet
  - telephone networks
  - transportation networks
- Biological networks
  - protein-protein interaction networks
  - gene regulation networks
  - the food web

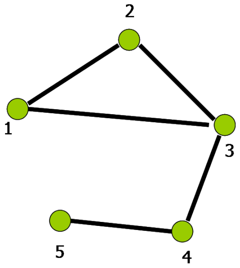
Foundational bibliography on complex networks

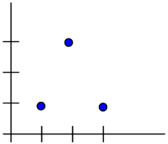
Around 1999...

- Watts and Strogatz, "Dynamics and small-world phenomenon"
- Faloutsos, Faloutsos and Faloutsos, "On power-law relationships of the Internet Topology"
- Kleinberg et al., "The Web as a graph"
- Barabasi and Albert, "The emergence of scaling in real networks"

Some basic definitions: degree distribution

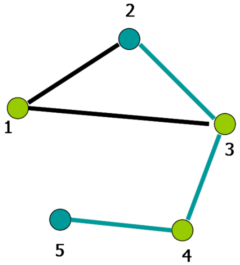
- degree $d(i)$ of vertex i: number of edges incident on i
- degree sequence: $[d(1), d(2), d(3), d(4), d(5)] = [2, 2, 3, 2, 1]$
- degree distribution: $[(1, 1), (2, 3), (3, 1)]$





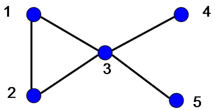
Some basic definitions: diameter

- diameter: the length of the longest shortest path between two vertices of the graph



Some basic definitions: clustering coefficient

- clustering coefficient of vertex i:
  - if $d(i) > 1$, is the number of edges between neighbors of i divided by $d(i)(d(i)-1)/2$
  - if $d(i) \leq 1$ can be defined as 0 or 1
- clustering coefficient of vertex 3: $1/6$
- clustering coefficient of vertex 1: 1

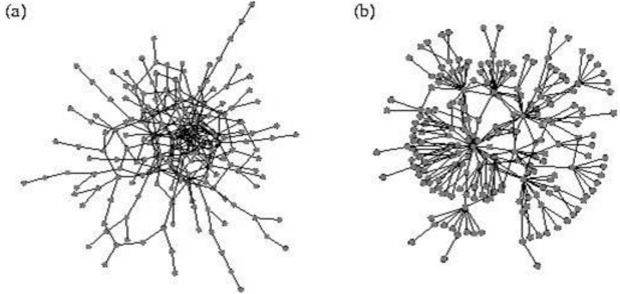


Characterization of complex networks

- Diameter, clustering coefficient, degree distribution.
- Betweenness centrality: number of short paths going through a vertex.
- Communities: can one identify cliques within the network?
- Correlations between degree and other quantities.
- Local motifs: What is the structure of the building blocks of complex networks?
    - Motifs: Subgraphs that have a significantly higher density in the observed network than in the randomizations of the same.
- Assortativity: do highly-connected nodes preferentially connect to other highly-connected nodes?

Assortativity

- A network is said to be assortatively mixed by degree if high degree vertices tend to connect to other high degree vertices.
- A network is disassortatively mixed by degree if high degree vertices tend to connect to low degree vertices.



Assortative and disassortative scale-free networks.

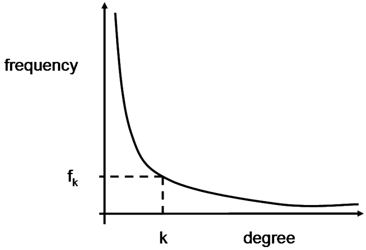
Real network properties

- Most vertices have only a small number of neighbors (degree), but there are some vertices with very high degree (power-law degree distribution)
  - scale-free networks
- If a vertex x is connected to y and z, then y and z are likely to be connected
  - high clustering coefficient
- Most vertices are just a few edges away on average.
  - small world networks
- Networks from very diverse areas (from internet to biological networks) have similar properties
  - Is it possible that there is a unifying underlying generative process?

Generating random graphs

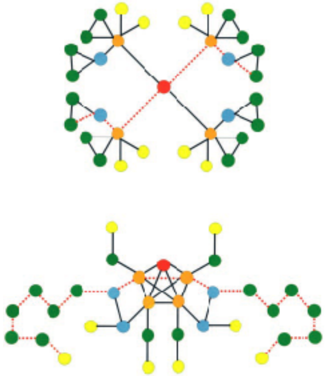
- Classic graph theory model (Erdős-Renyi)
  - each edge is generated independently with probability p
- Very well studied model but:
  - most vertices have about the same degree
  - the probability of two nodes being linked is independent of whether they share a neighbor
  - the average paths are short
- Real life networks are not "random" in this sense of randomness.
- Can we define a model that generates graphs with statistical properties similar to those in real life?

Degree distributions



- f_k = fraction of nodes with degree k = probability of a randomly selected node to have degree k
- Problem: find the probability distribution that best fits the observed data.

Degree distribution



These graphs have the same degree distribution but their diameter, modularity and robustness are very different.

Power-law distributions

- The degree distributions of most real-life networks follow a power law

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

  - there is a non-negligible fraction of nodes that has very high degree (hubs)
  - scale-free: no characteristic scale, average is not informative.
- In contrast with the random graph model!
  - Poisson degree distribution

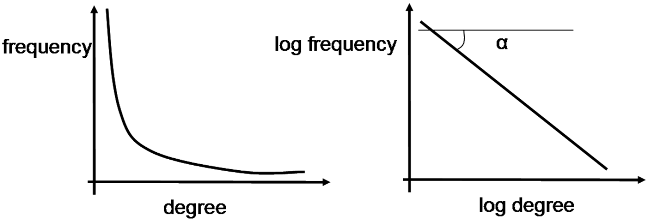
$$P(k) = \frac{(np)^k}{k!} e^{-np}$$

  - highly concentrated around the mean
  - the probability of very high degree nodes is exponentially small

Power-law signature

- Power-law distribution gives a line in the log-log plot

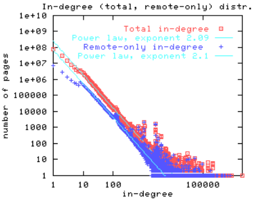
$$\log p(k) = -\alpha \log k + \log C$$

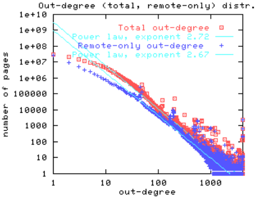


- α: power-law exponent (typically $2 \leq \alpha \leq 3$)

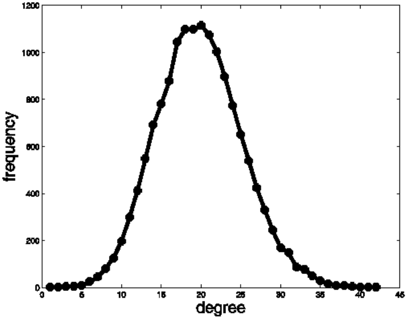
Example: the WWW graph

- In-degree distribution: Power-law distribution with exponent 2.1
- Out-degree distribution: Power-law distribution with exponent 2.7
- The fact that the exponent is greater than 2 implies that the expected value of the degree is a constant (not growing with n).
- Therefore, the expected number of edges is linear in the number of vertices n.





A random graph example



Expected degrees

- Average degree:
  - For random graphs: np.
  - For scale-free graphs, it is constant if $\alpha \geq 2$, and it diverges if $\alpha < 2$.
- Maximum degree:
  - For random graphs, the maximum degree is highly concentrated around the average degree.
  - For scale-free graphs $k_{\max} \approx n^{1/(\alpha-1)}$.

Connected components

- It is interesting to measure the size and distribution of the connected components, in particular, is there a giant component?
- Network Resilience: Study how the graph properties change when performing random or targeted node deletions.

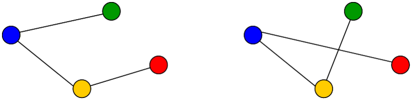
Motifs

- Most networks have the same characteristics with respect to global measurements... can we say something about the local structure of the networks?
- Motifs: Find small subgraphs that over-represented in the network.
- Finding interesting motifs: Count the frequency of the motifs of interest and compare against the frequency of the motif in a random graph with the same number of nodes and the same degree distribution.

Randomizing a network by edge swapping

Edge swapping (rewiring) algorithm: Randomly select and rewire two edges. Repeat many times.

This algorithm maintains the degree distribution. It is used to compare characteristic measured from a real network with those of randomized ones with the same degree distribution, for example, the presence of motifs.



What is a network model?

- Informally, a network model is a process (randomized or deterministic) for generating a graph
- Models of static graphs
  - input: a set of parameters Π, and the size of the graph n
  - output: a graph $G(\Pi, n)$
- Models of evolving graphs
  - input: a set of parameters Π, and an initial graph G_0
  - output: a graph G_t for each time t

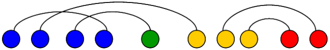
Graphs with given degree sequences

- The configuration model
  - input: the degree sequence $[d_1, d_2, \ldots, d_n]$
  - process:
    - Create d_i copies of vertex i
    - Take a random matching (pairing) of the copies, and then there will be one link from i to j for each link from a copy of i to a copy of j.
    - Self-loops and multiple edges are allowed.

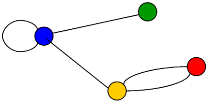
Example

Suppose that the degree sequence is

Create multiple copies of the nodes



Pair the nodes uniformly at random and generate the resulting network



Graphs with given expected degree sequences

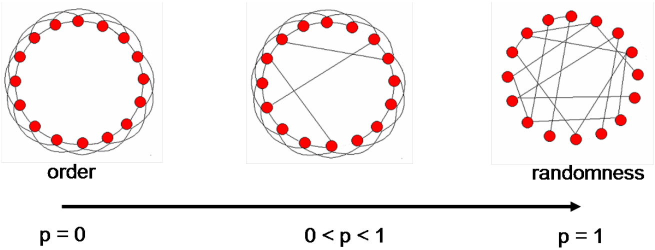
- input: the degree sequence $[d_1, d_2, \ldots, d_n]$ and the total number of edges m
- process: generate edge (i, j) with probability $d_i d_j / m$
- preserves the expected degrees
- easier to analyze.

Preferential Attachment in Networks

- First considered by Price (1965) as a model for citation networks.
  - each new paper is generated with m citations (mean)
  - new papers cite previous papers with probability proportional to their indegree (citations) plus one (to give some chance to papers with no citations).
- Power law with exponent $\alpha = 2 + 1/m$.
- The Barabasi-Albert model is similar and results in power law with exponent $\alpha = 3$.

Small World networks (Watts and Strogatz model, 1998)

- Start with a ring, where every vertex is connected to the next z vertices.
- With probability p, rewire two edges (or, add a shortcut to a uniformly chosen destination).



- For $0 < p < 1$, we have high clustering coefficient and small diameter.

Spread in networks

Understanding the spread of viruses (or rumors, information, failures etc) is one of the driving forces behind network analysis.

Percolation in networks

- Site Percolation: Each vertex of the network is randomly set as occupied or not-occupied. We are interested in measuring the size of the largest connected component of non-occupied vertices.
- Bond Percolation: Each edge of the network is randomly set as occupied or not-occupied. We are interested in measuring the size of the largest component of vertices connected by non-occupied edges.
- Good model for failures or attacks.

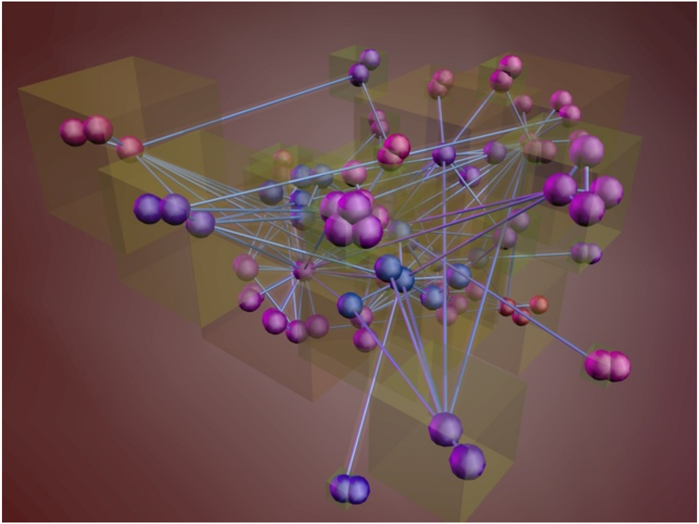
Percolation threshold

- How many vertices should be occupied in order for the network to not have a giant component? (the network does not percolate).
- For scale free graphs of power law exponent less than 3, there is always a giant component (the network always percolates).
- But... if the vertices are removed preferentially (according to degree), then it is easy to disconnect a scale free graph by removing a small fraction of the vertices.
- Scale-free graphs are resilient to random attacks, but sensitive to targeted attacks. For random networks there is smaller difference between the two.

Fractal dimension of scale-free networks

- Fractals look the same on all scales = 'scale-invariant'.
- In a recent work (C. Song, S. Havlin and H. A. Makse, 2005), the authors identify for some complex networks a power law relation between the number of boxes needed to cover the network and the size of the box, which defines a finite fractal dimension.
- A box of size k in a graph is a subset of vertices pairwise at distance at most k.
- We need the minimum number of boxes: NP-hard optimization problem! (clique covering in G^k). They use some heuristics.

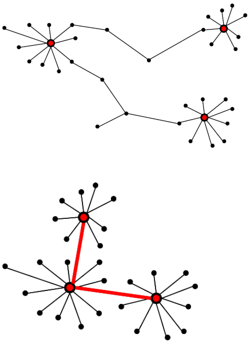
Box covering of a network



- Fractal networks:
  - WWW, biological networks.
  - Are characterized by the relation

$$N_B(\ell_B)/n \sim \ell_B^{-d_B}$$

    where d_B is the fractal dimension and $N_B(\ell_B)$ is the minimum number of boxes of size ℓ_B necessary to cover the network.
  - Are disassortative.
- Non-Fractal networks:
  - Internet, social networks (citations, IMDB), models based on uncorrelated preferential attachment.
  - Are assortative.

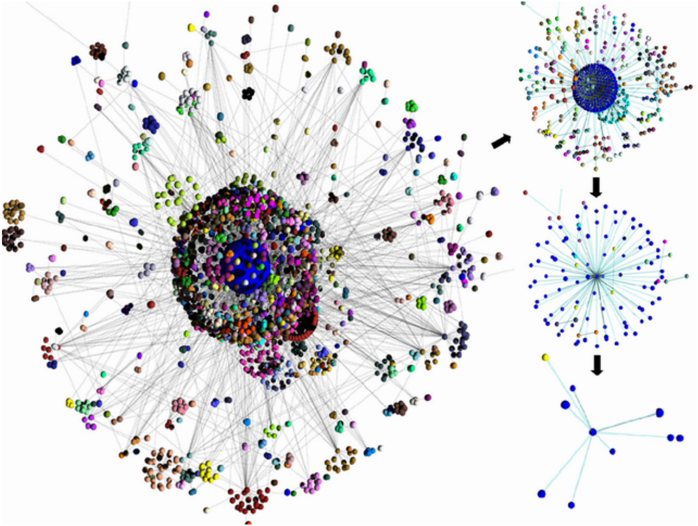


How to “zoom out” of a complex network?

Renormalization in Complex Networks: Now, regard each box as a single vertex and ask what is the degree distribution of the network of boxes at different scales ?

- The scale-free degree distribution is invariant under this renormalization.
- Internet is not fractal, but it is renormalizable.

Renormalization of the WWW



Acknowledgments: To Panajotis Tsaparas, Hernán Makse and Mariano Sigman, for some ideas for this talk.