Lecture X
Networks: an introduction

Francesco Campigli[1]

Computational Economics course
University of Florence, Italy

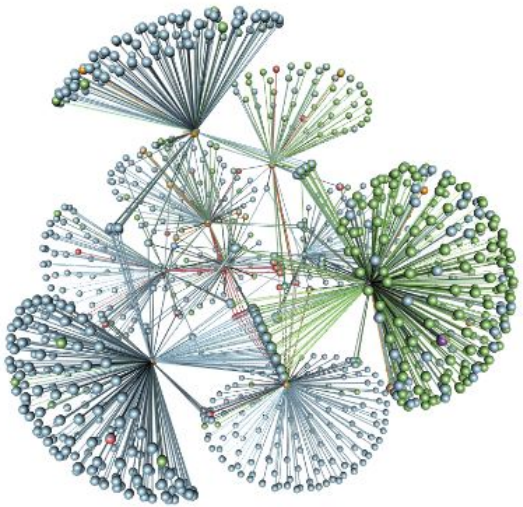
April 3, 2023

[1]Scuola Normale Superiore, Pisa, Italy.

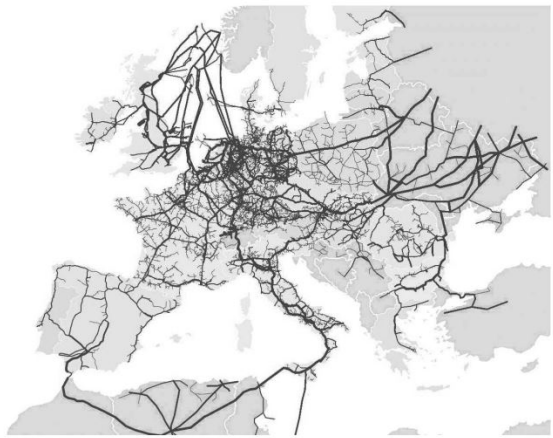
Contents

The basic idea

Network: a collection of points joined together in pairs by lines.

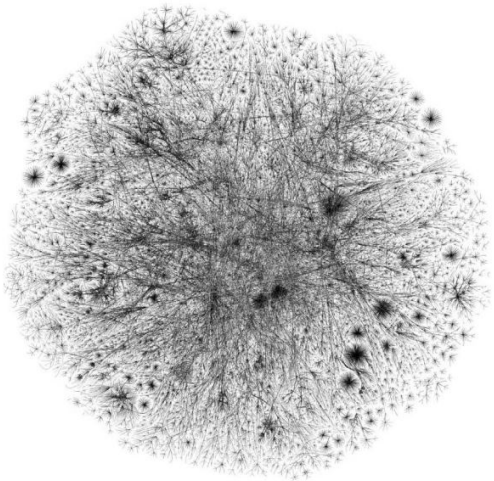


(a) HIV-1 protein

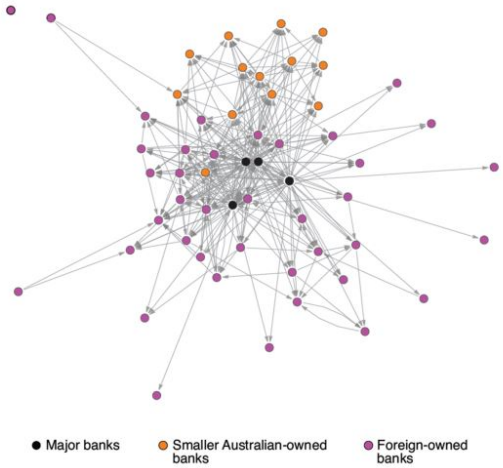


(b) Gas pipelines

The basic idea



(c) Internet



(d) Interbank

Why are so important?

Everything is a network.

Sources

- Main textbook: M.E.J. Newman, Networks: An Introduction. Oxford University Press, 2010
- Other sources: Barabasi, Vespignani, Dorogovtsev, Caldarelli...
- On Economic/Social networks: Jackson, Easley and Kleinberg, Vega-Redondo.
- Source for network data: http://math.nist.gov/RPozo/complex datasets.html (contains links to other repositories).
- Software packages: igraph (R), networkx (Python)
- Visualization software: Pajek, Gephi, Cytoscape...

Definition

A **network**, also called **graph** G, is a couple of sets V and E, and can be indicated as $G = (V, E)$, where:

- V is the set of NODES (or VERTICES)
- E is the set of LINKS (or ARCS)

Three elements:

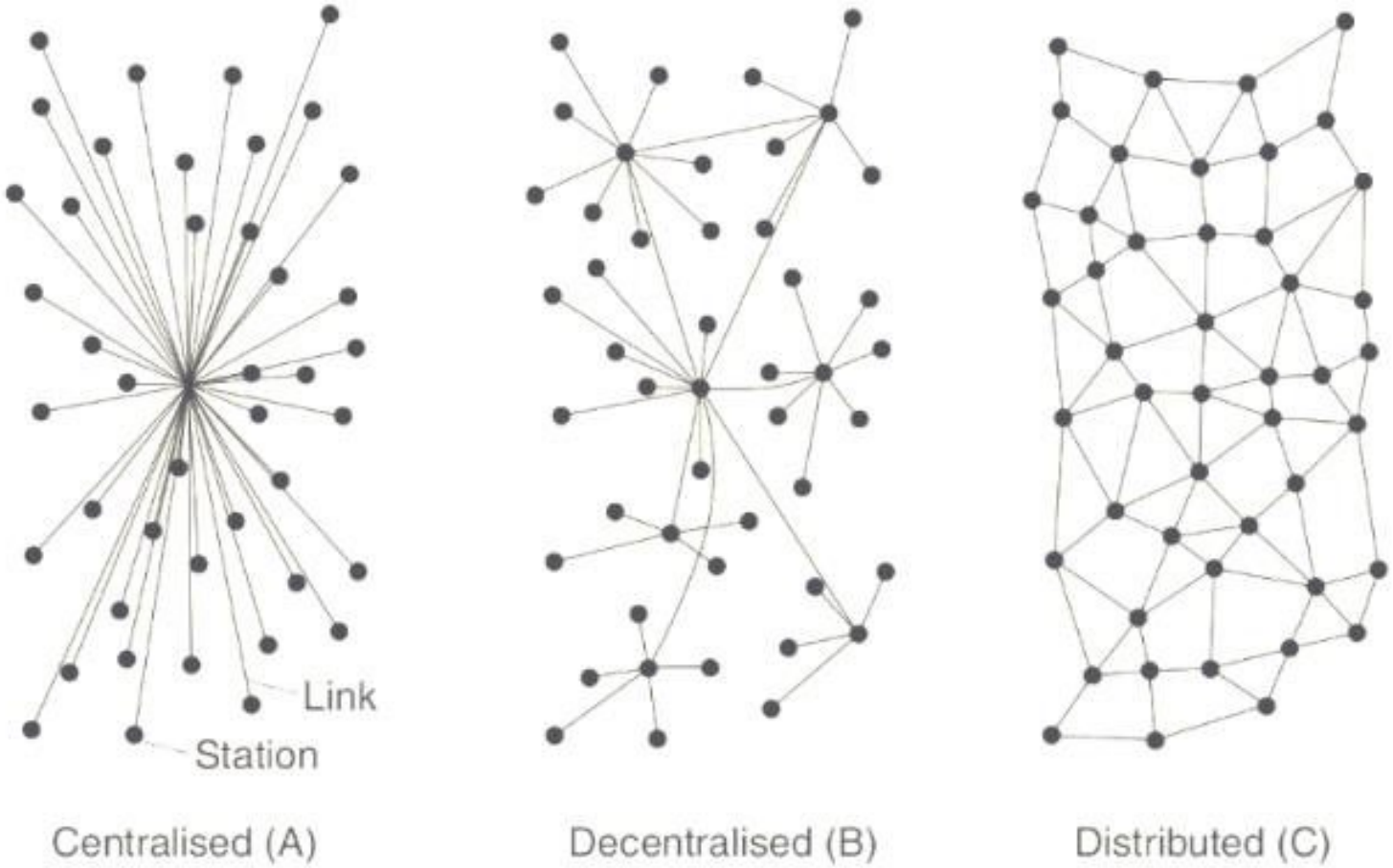
1. The graph is the **entire representation** of the network
2. Nodes are **objects that act** in the network
3. Edges are the **connections** between the nodes

The cardinality of V, $|V| = n$, is the number of vertices in the graph (order).

The cardinality of E, $|E| = m$, is the number of links in the graph (size).

Kind of Networks

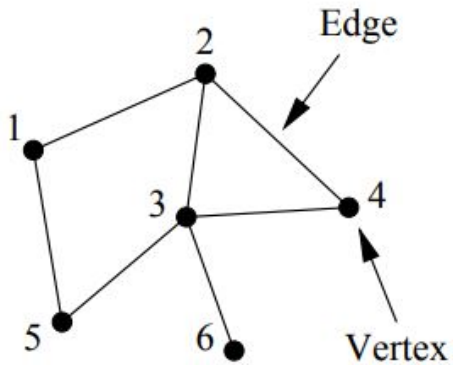
1. **Centralized networks** built around a single, **centralized** "master" node.
2. **Decentralized** network connections across multiple masters. Each of these separate nodes is a central unit that interacts independently with other nodes.
3. **Distributed** networks are composed of equal, interconnected nodes.



Adjacency matrix

- The structure of a network can be described by a n x n matrix: the (weighted) adjacency matrix $A = \{a_{ij}\}$.
- If two nodes i and j are not joined by a link, $a_{ij} = 0$, otherwise, $a_{ij} = 1$.

$$A = \begin{cases} 1, & \text{if there is an edge between vertices } i \text{ and } j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

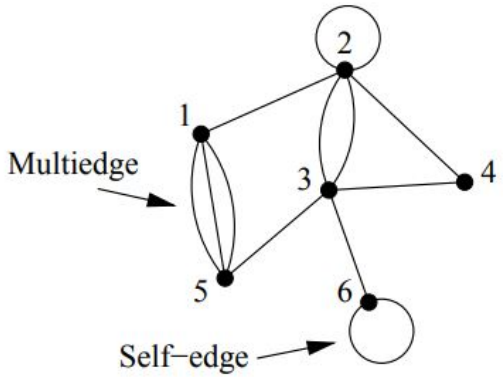


The degree of a node i is the number of its links: $k_i = \sum_j a_{ij}$.

Adjacency matrix

- There can be more than one edge between the same pair of vertices: **multiedges**.
- There can be no edges that connect vertices to themselves: **self-edges**.

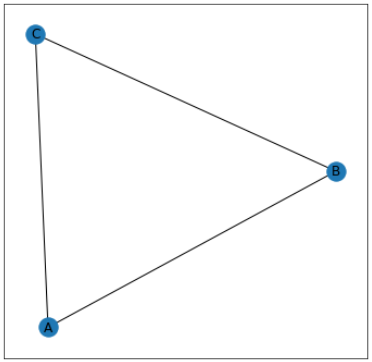
A network with neither self-edges nor multiedges is called a **simple network**.
A network with multiedges is called a **multi-graph**.



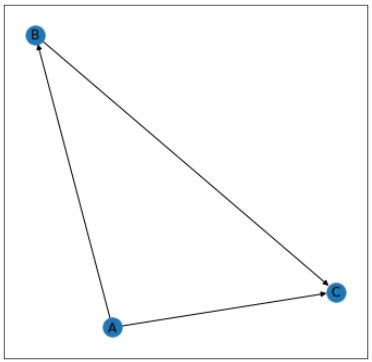
A multiedge is represented by setting the corresponding matrix element A_{ij} equal to the multiplicity of the edge.
A single self-edge from vertex i to itself is represented by setting the corresponding diagonal element A_{ii} of the matrix equal to 2.

Direct and Indirect Networks

1. If E is a set of non-ordered pairs of distinct elements in V, we have a **binary undirected network**. For undirected networks the adjacency matrix is symmetric, $A = A^T$.

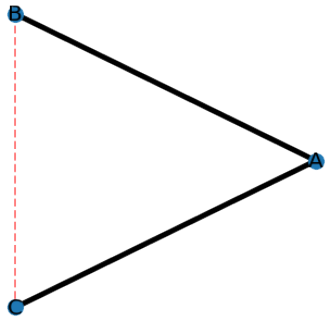


2. If E is a set of ordered pairs of distinct elements in V, we have a **binary directed network**. This means that a link connecting node i to j is different from a link connecting j to i.

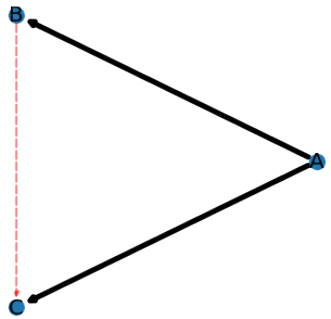


Direct and Indirect weighted Networks

1. If a binary undirected network is associated with a function $w : V \to \mathbb{R}^+$ we have a **weighted undirected network**: w_{ij} is the weight of the link $i - j$.
   If $w : V \to \mathbb{N}$, the weight can be seen as:
   - numbers quantifying the relationship between nodes.
   - a measure of how many single links are present between any two nodes.



2. If weights are associated with a directed network we have a **weighted directed network**.



Paths

- A **path** in a network is any sequence of vertices such that every consecutive pair of vertices in the sequence is connected by an edge in the network.
- The **length** of a path is the number of edges traversed along the path.
- The total number of paths of length 2 from node i to node j is:

$$N_{ij}^{(2)} = \sum_{k=1}^{n} a_{ik} a_{kj} = (A^2)_{ij}. \tag{2}$$

  and generalizing the number of paths $i \to j$ of length r is $(A^r)_{ij}$.
- The total number of **cycles** of length r is thus:

$$L_r = \sum_{i=1}^{n} (A^r)_{ii} = Tr[A^r]. \tag{3}$$

  What if the network is indirect? Proof: $L_r = \sum_i^n \phi_i$, where ϕ_i are the eigenvalues of A.

Network metrics

- The **distance** between two vertices i, j is the shortest number of edges to go from i to j.
- The **neighbors** of a vertex i are all vertices j which are connected to that vertex by a single edge ($d_{ij} = 1$):

$$d_{ij} = \min\{ \sum_{k, i \in P_{ij}} a_{ki} \}, \tag{4}$$

  where P_{ij} is a path connecting vertex i to vertex j.
- The **diameter** of a graph is given by the maximum of all distances between pairs.

Network metrics

- The **size**: $L = \sum_{i<j} a_{ij}$ (undirected) and $L = \sum_{i,j} a_{ij}$ (directed).
- The **connectance** is a measure of the density of links in the network:
  - Undirected network: $c^{(und)} = \frac{2L}{n(n-1)}$
  - Directed networks: $c^{(dir)} = \frac{L}{n(n-1)}$
- The **degree** of a node is the number of neighbors:
  - Undirected network: $k_i = \sum_j a_{ij}$.
  - Directed networks: indegree $k_i^{in} = \sum_j a_{ji}$, outdegree $k_i^{out} = \sum_j a_{ij}$, total $= k_i^{tot} = k_i^{in} + k_i^{out}$.
- For a weighted networks the **strength** of a node is:

$$s_i = \sum_j w_{ij}. \tag{5}$$

- Similarly one defines the in-strength and the out-strength

Centrality Measures

- Broadly speaking, the centrality of a node (or of an edge) of a network is a measure of the importance of the node:
  - how influential is a person in a social network
  - how critical is an element in an infrastructure network,
  - what is the disease-spreading capacity of an individual
  - what is the most systemically important financial institution
- Loose definition: many metrics!
  - Degree centrality
  - Betweenness centrality
  - Closeness centrality
  - Eigenvector centrality

Centrality Measures

- **Degree centrality**. The degree of a node is an obvious measure of centrality. Only local information:

$$dc_i = \frac{k_i}{N-1}, \tag{6}$$

where N is the total number of nodes.

- **Betweenness centrality** of a node v is computed in the following way: for each pair of vertices (i, j) one identifies the N_{ij} shortest paths between them and computes the number $N_{ij}(v)$ of them that pass through v. The betweenness of node v is:

$$C_B(v) = \frac{1}{(n-1)(n-2)} \sum_{i,j \neq v} \frac{N_{ij}(v)}{N_{ij}} \tag{7}$$

i.e. the average fraction of shortest paths passing through v, where the average is taken across all the pairs of vertices.

Centrality Measures

- A geodesic path is a path between two vertices such that no shortest path exists.
  - Geodesic paths are necessarily self-avoiding.
  - Geodesic paths are not necessarily unique.
  - However the distance d_{ij} (possibly infinite if no path exists) is unique.
- The diameter of a graph is the length of the longest geodesic path between any pair of vertices (for which such path exists).
- **Closeness centrality** of node i is the inverse of the mean geodesic distance of i from all other nodes:

$$C_i = \frac{n}{\sum_j d_{ij}}. \tag{8}$$

Notice: sometimes $n-1$ is used for the normalization and as an alternative the harmonic mean is used.

Centrality Measures

A vertex's importance in a network is increased by having connections to other vertices that are themselves important.

- Consider an undirected network. Let start guessing $x_i = 1, \forall i$. New centrality estimate is **x'** $= A\mathbf{x}$ and iterating $\mathbf{x}(t) = A^t\mathbf{x}(0)$.
- By decomposing $\mathbf{x}(0)$ in terms of the eigenvectors of A, $\mathbf{x}(0) = \sum_i c_i\mathbf{v}_i$, it is:

$$\mathbf{x}(t) = A^t \sum_i c_i\mathbf{v}_i = \sum_i c_i k_i^t \mathbf{v}_i = k_1^t \sum_i c_i \left[\frac{k_i}{k_1}\right]^t \mathbf{v}_i, \tag{9}$$

  where k_1 is the largest eigenvalue. **Notice:** for $t \to \infty$ only the contribution of the first eigenvalue survives, hence $A\mathbf{x} = k_1\mathbf{x}$.
- **Eigenvector centrality.** Eigenvector centrality is simply defined in terms of the adjacency matrix $A = \{a_{ij}\}$, where a_{ii} can be either a binary or a non negative realvalue (weighted matrix). The vector $\mathbf{x} = (x_1, \ldots, x_n)'$ satisfies:

$$A\mathbf{x} = k_1\mathbf{x}, \tag{10}$$

  i.e. it is a right eigenvector of the adjacency matrix corresponding to the maximal eigenvalue k_1.

Centrality Measures

The eigenvector centrality measures the importance of a node based on the score of its neighbors. Contrary to the previous measures, this one is not based on distance among nodes, but depends recursively on the centrality of the neighbors. In vector notation, the eigenvector centralityc is the vector that solves the equation $Wc = \lambda c$ where λ is the largest eigenvalue. In terms of recursive expression we can define the eigenvector centrality of node i as:

$$c_i = \frac{1}{\lambda} \sum_j W_{ij} c_j \tag{11}$$

It is intrinsically based on the spectral properties of adjacency matrix. So it provides a different approach to assess node centrality.

Graph Laplacian

- Consider the diffusion on a network: ψ_i is the amount of "substance" at node i and its dynamics is described by

$$\frac{d\psi_i}{dt} = C \sum_j a_{ij} \left(\psi_j - \psi_i\right)$$

where C is a diffusion constant.

- This can be rewritten as

$$\frac{d\psi_i}{dt} = C \sum_j a_{ij}\psi_j - C\psi_i \sum_j a_{ij} = C \sum_j \left(a_{ij} - \delta_{ij}k_i\right)\psi_j$$

or in matrix notation

$$\frac{d\vec{\psi}}{dt} = C(A - D)\vec{\psi}$$

where D is the diagonal matrix with $D_{ii} = k_i$.

Graph Laplacian

- The graph Laplacian is the matrix is $L = D - A$
- The diffusion equation becomes

$$\frac{d\vec{\psi}}{dt} + CL\vec{\psi} = 0$$

similar to $\partial_t \vec{\psi} + C\nabla^2 \vec{\psi} = 0$.

- The solution of the diffusion equation can be found by setting

$$\vec{\psi}(t) = \sum_i a_i(t)\vec{v}_i$$

where $\vec{v}_i$ are the eigenvectors of L associated to the eigenvalues λ_i. Substituting in the diffusion equation we get

$$\dot{a}_i + C\lambda_i a_i = 0$$

with solution

$$a_i(t) = a_i(0)e^{-C\lambda_i t}$$

Graph Laplacian

- It can be shown (see Newman) that $\lambda_i \geq 0 \forall i$, thus the solution of the diffusion equation tends to an equilibrium value when $t \to \infty$.
- Note that at least one eigenvalue of L is zero. In fact the vector $(1, 1, \ldots)^T$ is the associated eigenvector

$$\sum_j L_{ij} \times 1 = \sum_j (\delta_{ij} k_j - a_{ij}) = k_i - \sum_j a_{ij} = 0$$

- If the network is divided in c disconnected components, the adjacency matrix and the Laplacian are block diagonal and any vector of ones in all the elements on one component and zero in the elements of other components will be an eigenvectors with eigenvalue zero
- Hence the second largest eigenvalue of the graph Laplacian is non-zero iff the network is connected. This eigenvalue is called algebraic connectivity or spectral gap of the network.

Application: mean first passage time

- Consider a random walker on a network and let $p_i(t)$ the probability that the walker is at node i at (discrete) time t. The evolution equation for $\vec{p}$ is

$$p_i(t) = \sum_j \frac{a_{ij}}{k_j} p_i(t-1)$$

or $\vec{p}(t) = AD^{-1}\vec{p}(t-1)$.

- In the limit $t \to \infty$ the stationary probability satisfies $\vec{p} = AD^{-1}\vec{p}$ or

$$\left(I - AD^{-1}\right)\vec{p} = (D-A)D^{-1}\vec{p} = LD^{-1}\vec{p} = 0$$

i.e. $D^{-1}\vec{p}$ is an eigenvector of the Laplacian with eigenvalue 0 .

- If the network has only one component, there is only one eigenvector $(1, 1, \ldots)^T$ with eigenvalue 0 . Hence $\vec{p} \propto D\vec{1}$, i.e. $p_i \propto k_i$. Hence the stationary probability of the random walker is proportional to the degree.

Application: mean first passage time

- The first passage time for a random walk from a vertex u to a vertex v is the number of steps before a walk starting at u reaches v.
- First passage time is a random variable, we are interested here to the mean value.
- We modify the walk to make it an absorbing random walk, i.e. if the walk reaches v it remains there forever.
- Let $p_v(t)$ be the probability that the walker is at v (i.e. absorbed) at time t.
- This is also the probability that the walk has a first passage time to v that is less than or equal to t.
- Hence the mean first passage time can be rewritten as

$$\tau_v = \sum_{t=0}^{\infty} t \left[p_v(t) - p_v(t-1) \right]$$

Application: mean first passage time

- $\forall i \neq v$

$$p_i(t) = \sum_j \frac{a_{ij}}{k_j} p_j(t-1) = \sum_{j \neq v} \frac{a_{ij}}{k_j} p_j(t-1)$$

because $a_{iv} = 0$ (absorbing state). But if $i \neq v$ then there are no terms in A_{vj} in the sum either thus we can write

$$\vec{p}'(t) = A' D^{-1} \vec{p}'(t-1) = \left[A' D'^{-1}\right]^t \vec{p}'(0)$$

where' means that we have removed the v-th row and column.

- Clearly $p_v(t) = 1 - \overrightarrow{1}^\top \vec{p}'(t)$, thus

$$\tau_v = \sum_{t=0}^{\infty} t \, \overrightarrow{1}^T \left[\vec{p}'(t-1) - \vec{p}'(t)\right] = \overrightarrow{1}^T \left[I - A' D^{-1}\right]^{-1} \vec{p}'(0)$$

where we have used

$$\sum_{t=0}^{\infty} t \left[M^{t-1} - M^t\right] = \left[I - M\right]^{-1}$$

Application: mean first passage time

- Since

$$\left[I - A' D'^{-1}\right]^{-1} = D' \left[D' - A'\right]^{-1} = D' L'^{-1}$$

we finally obtain

$$\tau_v = \overrightarrow{1}^{\top} D' L'^{-1} \vec{p}'(0)$$

- L' is the graph Laplacian where the v-th row and column are removed and is called reduced Laplacian.
- Note that even if L is not invertible (because it has a zero eigenvalue), L' can have an inverse because $\overrightarrow{1}$ is not in general an eigenvector of L'.