Lecture 1: Introduction (networks, small worlds, scale-free property, clustering)

Prof. Sotiris Nikoletseas

University of Patras
and CTI

ΥΔΑ ΜΔΕ, Patras
2020 - 2021

Real Networks

- Networks can be abstracted by nodes connected by links
- Networks emerge in technology (Internet, WWW), society (transportation networks, electricity networks, social networks), biology (cells, proteins)
- Many real networks are neither purely regular nor purely random
- They exhibit complex topological features, with properties like:
  - small-world (short path lengths among any two nodes)
  - hubs (power-law degree distributions)
  - high clustering of nodes

Networks (graphs)

- a network (or graph) is a set of nodes (vertices) connected by links (edges).

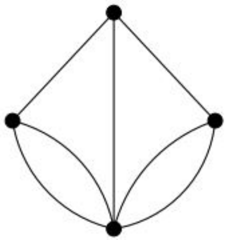


Figure: *A graph*

Networks (graphs): Basic Definitions

- **degree of a node:** The number of its connections (adjacent nodes).
- **regular graph:** all nodes have same degree.
- **simple graph:** no multiple links among any two nodes, neither any loops of length 1 (nodes connected to themselves).
- **multigraph:** a graph with multiple links or self-loops.
- **path:** a sequence of adjacent nodes and links with no repeated nodes.
- **cycle (or loop):** a closed path where only the start and end nodes coincide.
- **tree:** a connected graph without loops (trees are simpler to analyse).
- **network size:** the number of nodes N. Also, let L the number of links. In trees, it is $L = N - 1$.

Networks (graphs): Abstraction (I)

Networks represent a wide variety of structures in complex systems comprised of entities (nodes) and their relations (modelled by edges):

- edges representing tangible, physical links, such as cables, roads, electricity supply, neural networks.
- edges representing physical interactions, such as biological interactions of proteins.
- edges representing etheral, intangible connections irrespective of the physical layer, such as relations of pages in the web or in a network of airports.

Networks (graphs): Abstraction (II)

- edges representing geographic closeness, such as among countries in a map, sensors in an IoT deployment, cells connected in tissues.
- edges representing social connections, such as friendship, collaboration, common interests etc.
- edges representing conceptual linking, such as in dictionaries and citation networks.

Networks (graphs): An airline network

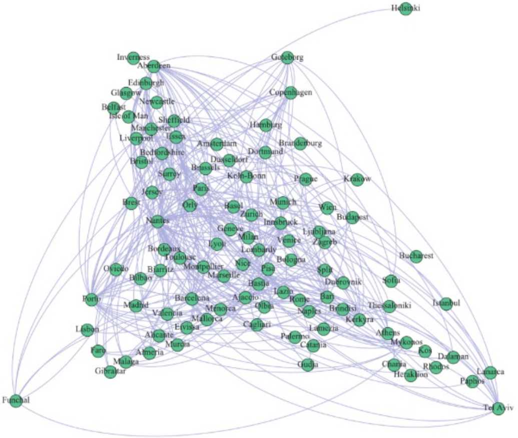


Figure: *An airline transportation network in Europe*

Networks (graphs): A city network

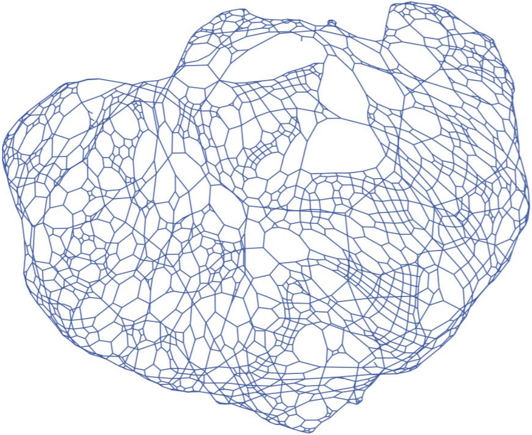


Figure: *An urban street network*

Networks (graphs): A conceptual network

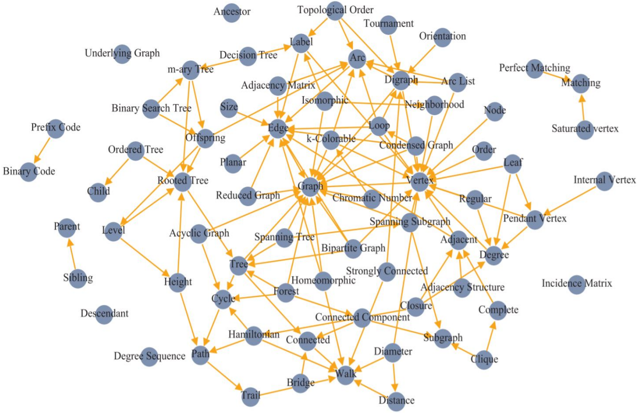


Figure: *Relational network of concepts in network theory*

Examples of graphs (I)

- a complete graph: all nodes connected to each other $\Rightarrow L = \binom{N}{2}$.
- a star: a "central" node to which all rest nodes are linked $\Rightarrow$ maximum separation among nodes is 2 (it is the most compact tree).
- comb and brush graphs.

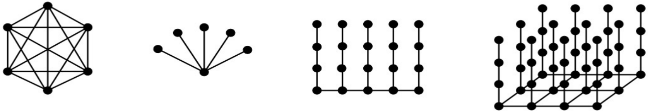


Figure: *Examples of different graphs*

Examples of graphs (I)

- (q, g)-cage graphs: a regular graph with minimum number of links such that degrees are q and a given length g of the shortest cycle. e.g. the Petersen graph is a (3,5)-cage graph.
- hypergraph: generalized graphs where links (edges) are subsets of nodes (a graph is a hypergraph where all edges have cardinality 2).

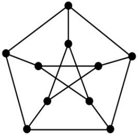
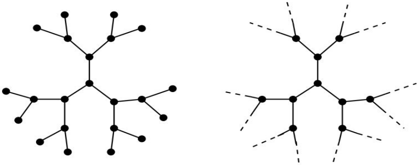


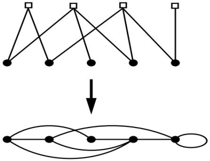
Figure: *The Petersen graph*

Examples of graphs (III)

- a Cayley tree: a regular tree with a central node (root) and a boundary with a finite fraction of dead end nodes (leaves).
- a Bethe lattice: an extended, infinite Cayley tree without a boundary and any dead ends $\Rightarrow$ all nodes in a Bethe lattice are equivalent, there is no central node.
- bipartite graphs: nodes partitioned in two sets with links only among different parts (e.g. networks of nodes-authors and nodes-papers modeling scientific co-authorship).



Cayley tree and Bethe lattice



Bipartite graph

Shortest path length

- distance ℓ_{ij} : the length (number of links) of the shortest path between nodes i and j
- Two notions of node separation:
  - mean internode distance $\bar{\ell}$: the average of the ℓ_{ij} distances over all (i, j) node pairs (for which $\ell_{ij} < \infty$)
  - diameter ℓ_D : the maximum of the various ℓ_{ij} ($i.e.$, the maximum internode distance)
- the dependence of $\bar{\ell}$ or ℓ_D on network size N is characteristic in different network types:
  - in "compact" networks, $\bar{\ell}(N)$ grows slower with N
  - in "looser" structure (like lattices) the growth of $\bar{\ell}(N)$ is fast

Lattices vs Trees (I)

- in finite lattices, the size dependence $\bar{\ell}(N)$ is power-law:

$$\bar{\ell} \sim N^{1/d},$$

where d is the dimension of the lattice (an integer).
$e.g.$ in a regular finite lattice of dimension 2, it is obviously $\bar{\ell} \sim \sqrt{N}$ $(= N^{1/2})$

In other words, such networks are "large worlds" with large distances, $e.g.$ when $N = 10^{12}$ then $\bar{\ell} \sim 10^6$.

Lattices vs Trees (II)

- in contrast, trees are “small” worlds: consider the q-regular Bethe lattice and Cayley tree. Let $b = q - 1$ the “branching” constant. Then, the number of nodes $n = n(\ell)$ within distance ℓ from the root is:

$$n = 1 + q + qb + qb^2 + .. + qb^{\ell-1} =$$

$$= 1 + q(1 + b + b^2 + ... + b^{\ell-1}) = 1 + q\frac{b^\ell - 1}{b - 1}$$

$$n \sim b^\ell \Rightarrow \bar{\ell} \sim \frac{\ln N}{\ln b},$$

which grows much slower than the power-law in lattices. As an example, if a Cayley tree has 10^{12} nodes of degree 5, then $\bar{\ell} \sim 10$, which is dramatically smaller than in the lattice case.

Milgram's experiment - six degrees of separation (I)

- the small-world phenomenon first observed in a social network of acquaintances by the social psychologist Stanley Milgram (1967).
- question: how many intermediate social links separate two randomly chosen, remotely located individuals?
- experiment: choose a random person in Omaha, Nebraska and a random person in Boston. The Omaha person should either send a letter to the Boston person, or to another person who may know the Boston person and so on. Repeat this for a very large number of Omaha-Boston person-pairs.

Milgram's experiment - six degrees of separation (II)

- counter-intuitive finding: an essential number of letters made it to the target, after passing on average only 5.5 intermediate persons
  ⇒"six degrees of separation"

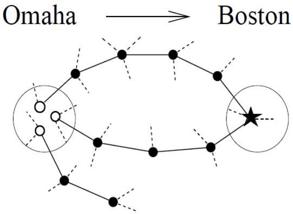


Figure: *How Stanley Milgram scanned a net of acquaintances in the US. Notice that some chains of acquaintances were broken off*

Directed networks

- they include directed connections
- example: networks of citations in scientific papers
  - nodes: papers
  - links: directed from a paper to papers cited by the first paper (which was published later in time)

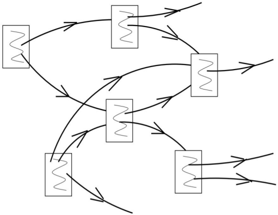


Figure: *Network of citations in scientific papers*

Random networks (I)

- a random network is not a single graph, but a statistical ensemble (a probability distribution) of graphs. This ensemble is created via a random experiment and each member of the resulting ensemble has a certain statistical weight (probability).
- example: the $G_{n,p}$ random graph ("Gilbert" model)
  - n labelled nodes (i = 1,2, ..., n)
  - each link exists with some probability p, independently for the various links

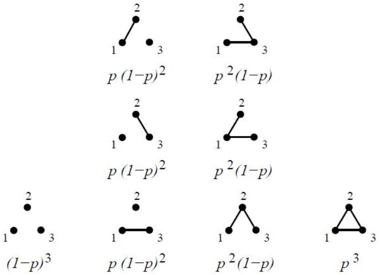


Figure: *The Gilbert model of a random graph (the $G_{n,p}$ model) for n=3*

Random networks (II)

- **random geometric graphs ($G_{n,R}$)**
  - throw randomly, uniformly n points in $[0,1]^2$.
  - connect any two points by a link iff their Euclidean distance is at most R

  (motivation: sensor networks)
- **random neighbor graphs ($G_{n,k}$)**
  - throw randomly, uniformly n points in $[0,1]^2$.
  - connect each node to each k nearest neighbors.

  (motivation: power control in wireless networks)
- **random intersection graphs ($G_{n,m,p}$)**
  - n nodes and m labels
  - each node randomly, independently chooses labels with probability p.
  - two nodes are linked iff they share at least one label.

  (motivation: social networks, frequency assignment, cryptography in wireless networks)

Degree distribution (I)

- the degree distribution $P(q)$ is the probability that a randomly chosen node in a random network has degree q:

$$P(q) = \frac{\langle N(q) \rangle}{N},$$

  where $\langle N(q) \rangle$ is the average number of nodes of degree q in the network, where averaging is taken over the whole statistical ensemble.
- Clearly $N = \sum_q \langle N(q) \rangle$.
- Empirically, in a single graph g of the ensemble, we observe the number $N_g(q)$ of nodes of degree q. The ratio $P_g(q) = \frac{N_g(q)}{N}$ is also usually called degree distribution and in the infinite network limit, it coincides with $P(q)$.

Degree distribution (II)

- the degree distribution is the simplest statistic of a random network, however in many cases the insight it offers is very informative.
- all parts of the degree distribution (including the low- and high-degree parts) are important.
- particularly, the decay rate of the degree distribution is very characteristic:
  - in $G_{n,p}$ random graphs, degree distribution decays quite fast:

    $$P(q) \sim \frac{1}{q!} \quad \text{(for large } q\text{)}$$

    and practically there are no strongly connected hubs
  - in contrast, in several real world networks (like the Internet, cellular nets) degree distributions decay much slower, and hubs of essential role occur

    $$P(q) \sim q^{-\gamma} \quad \text{(for large } q\text{)}$$

    where γ constant $\Rightarrow$ power-law decay rate

Degree distribution (III)

Most common degree distributions in complex networks:

- Poisson distribution: in random networks of n nodes and independent edge probability p, it is:

$$Pr\{deg(u) = k\} = \binom{n-1}{k} p^k (1-p)^{n-1-k} \ (binomial)$$

which for large n (and mean degree $np = \bar{k}$ constant) tends to the Poisson distribution:

$$Pr\{deg(u) = k\} \rightarrow \frac{\bar{k}^k \cdot e^{-\bar{k}}}{k!}$$

(which becomes the normal distribution for large k)

Degree distribution (IV)

- exponential distribution: in emerging networks starting with a single node and nodes added one at a time attaching themselves randomly to existing nodes (thus, the newer the node the lower its expected degree is):

$$Pr\{deg(u) = k\} = A \cdot e^{-\frac{k}{\bar{k}}}$$

- power-law distribution, in a network where new nodes added attach themselves preferentially to nodes with high degree it is:

$$Pr\{deg(v) = k\} = B \cdot k^{-\gamma}$$

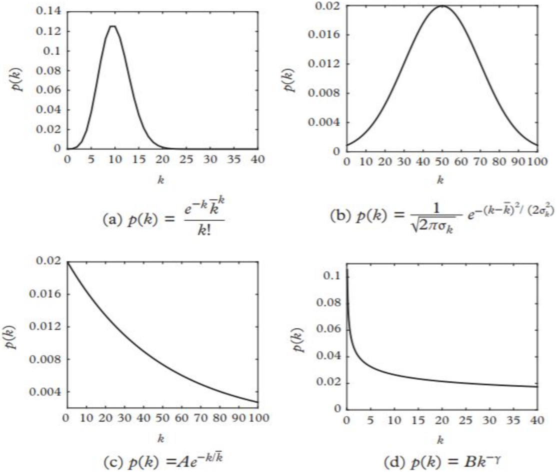


Figure: *Common examples of degree distributions found in complex networks*

The scale-free property

- This power-law degree distribution is also called scale-free and networks with such scale-free degrees are called scale free networks.
- The “scale-free” notion actually implies the absence of a typical node degree in the network; the network includes several strong hubs, in contrast to the rather regular structure of random networks.
- Strictly, the term “scale-free” refers to the fact that a power-law distribution $q^{-\gamma}$ has the following property:
  a re-scaling of the degree q by a constant c to cq $(c \to cq)$ only has a constant multiplicate effect to $(cq)^{-\gamma} = c^{-\gamma} \cdot q^{-\gamma}$. This makes few nodes of very high degree quite probable, in contrast to random networks where strong hubs are highly improbable.

Clustering (I)

- It is about how the nearest neighbors of a node are interconnected, so it is a non-local characteristic, going one step further than degree.
- The clustering coefficient of a node is the probability that two nearest neighbors of it are neighbors themselves. Thus, if node j has q_j nearest neighbors (degree q_j) and there are t_j connections among these q_j neighbors then the clustering coefficient of j is:

$$c_j = \frac{t_j}{q_j(q_j - 1)/2}$$

  since there are $\binom{q_j}{2}$ potential connections. This is the Watt-Strogatz clustering coefficient.
- When all neighbors are connected to each other, then $c_j = 1$.
- When there are no connections among them, then $c_j = 0$.
- Most real networks have strong clustering.

Clustering (II)

- Actually, clustering relates to the presence of many triangles in a network; such a feature generally results from high transitivity e.g. in a social network if Bob and Phil are both friends of Joe then they may probably meet and become friends to each other, forming a triangle.
- In this respect, note that t_j equals the number of triangles (loops of length 3) attached to a node.

Clustering (III)

- The average clustering coefficient in a fixed network of n nodes is:

$$\bar{c} = \frac{1}{n} \sum_j c_j$$

- The expected clustering coefficient of a node of degree q in a random network is:

$$\bar{c}(q) = \langle c_j(q) \rangle$$

i.e. we take the average over all nodes of degree q.

- The mean clustering coefficient of the whole network is:

$$\bar{c} = \sum_q P(q)\bar{c}(q)$$

where $P(q)$ the degree distribution.

The Newman Clustering coefficient (I)

- It is another way to quantify global clustering in a network (also called transitivity index). It is

$$c = \frac{3 \cdot t}{|P_2|},$$

where t the total number of triangles and $|P_2|$ the total number of paths of length 2 in the network (the latter counts all potential three-way relationships, i.e. the number of all connected triples of a node and two of its nearest neighbors).

The Newman Clustering coefficient (II)

- In general, the Watts-Strogatz coefficient focuses on local clustering while the Newman coefficient focuses on how the network is clustered as a whole. In real world networks, the two indices often correlate well:

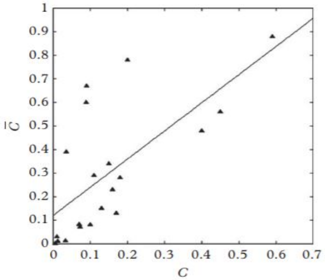


Figure: *Correlation between Watts-Strogatz ($\bar{C}$) and Newman (C) clustering coefficients for 20 real-world networks*

Course Themes

- Introduction (networks, small worlds, scale-free property, clustering)
- Classical random graphs (loops, diameter, connectivity)
- Small worlds, scale-free networks, generating networks of arbitrary degree
- Random Networks I - The method of positive probability
- Random Networks II - Linearity of Expectation
- The Second Moment Method
- Randomized Algorithms
- How the scale-free property emerges: preferential attachment
- Spreading phenomena/how epidemics emerge