Introduction to network metrics

Ramon Ferrer-i-Cancho & Argimiro Arratia

Universitat Politècnica de Catalunya

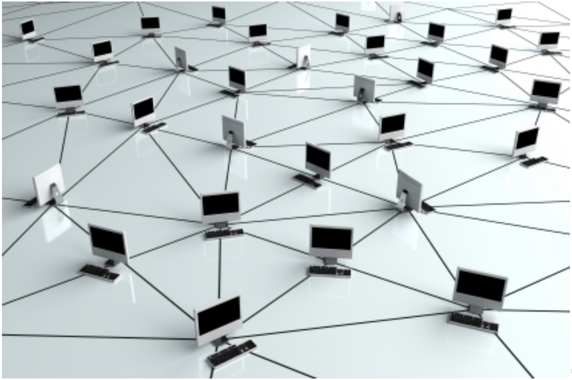
Version 0.4
Complex and Social Networks (2020-2021)
Master in Innovation and Research in Informatics (MIRI)

Official website: www.cs.upc.edu/~csn/
Contact:

- Ramon Ferrer-i-Cancho, rferrericancho@cs.upc.edu, http://www.cs.upc.edu/~rferrericancho/
- Argimiro Arratia, argimiro@cs.upc.edu, http://www.cs.upc.edu/~argimiro/

Network analysis

Two major approaches: visual and statistical analysis (e.g., large scale properties).



(from Webopedia)

Statistical analysis: compression of information (e.g., one value that summarizes some aspect of the network).

Perspectives

Metrics as compression of an adjacency matrix.
Three perspectives:

- Distance between nodes.
- Transitivity
- Mixing (properties of vertices making an edge).

Geodesic path

- Geodesic path between two vertices u and v = shortest path between u and v [Newman, 2010]
- d_{ij}: length of a geodesic path from the i-th to the j-th vertex (network or topological distance between i and j).
-
  - $d_{ij} = 1$ if i and j are connected.
  - $d_{ij} = \infty$ if i and j are in different **connected components**.
- Computed with a breadth-first search algorithm (in unweighted undirected networks).

Local distance measures

l_i: mean geodesic distance from vertex i

- Definitions:

$$l_i = \frac{1}{N}\sum_{j=1}^{N} d_{ij} \quad \text{or}$$

$$l_i = \frac{1}{N-1}\sum_{j=1(i\neq j)}^{N} d_{ij} \quad \text{as } d_{ii} = 0$$

C_i: closeness centrality of vertex i.

- Definition (harmonic mean)

$$C_i = \frac{1}{N-1}\sum_{j=1(i\neq j)}^{N} \frac{1}{d_{ij}},$$

as $d_{ii} = 0$.

- Better than $C_i' = 1/l_i$.

Global distance metrics

- Diameter: largest geodesic distance.
- Mean (geodesic distance):

$$l = \frac{1}{N} \sum_{i=1}^{N} l_i$$

  - Problem: l might be ∞.
  - Solutions: focus on the largest connected component, mean over l within each connected component, ...
- Mean closeness centrality:

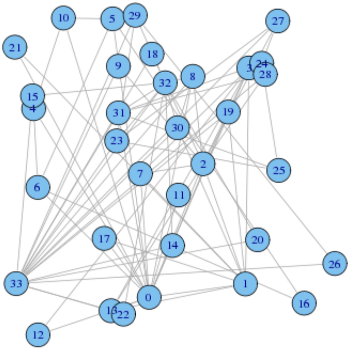
$$C = \frac{1}{N} \sum_{i=1}^{N} C_i$$

Global distance metrics

- Closeness measures have rarely been used (for historical reasons).
- The closeness centrality of a vertex can be seen as measure of the importance of a vertex (alternative approaches: degree, PageRank,...).

Transitivity

Zachary's Karate Club



- A relation $\circ$ is transitive if $a \circ b$ and $b \circ c$ imply $a \circ c$.
- Example: $a \circ b = a$ and b are friends.
- Edges as relations.
- Perfect transitivity: clique (complete graph) but real network are not cliques.
- Big question: how transitive are (social) networks?

Clustering coefficient

- A path of length two uvw is closed if u and w are connected.
$$C = \frac{\text{number of closed paths of length 2}}{\text{number of paths of length 2}}$$
A proportion of transitive triples
- $C = 1$ perfect transitivity / $C = 0$ no transitivity (**e.g.,: ?**).
- Algorithm: Consider each vertex as v in the path uvw, checking if u and w are connected (only vertices of degree ≥ 2 matter).
- Number of paths of length $2 = ?$.
- Equivalently:
$$C = \frac{\text{number of triangles} \times 3}{\text{number of connected triples of vertices}}$$
- Key: triangle = set of three nodes forming a clique; number of connected triples = number of labelled trees of 3 vertices

Alternative clustering coefficient

Watts & Strogatz (WS) clustering coefficient
[Watts and Strogatz, 1998]

- Local clustering:

$$C_i = \frac{\text{number of pairs of neighbors of } i \text{ that are connected}}{\text{number of pairs of neighbours of } i}$$

- Assuming undirected graph without loops:

$$C_i = \frac{\sum_{j=1}^{N} \sum_{k=1}^{j-1} a_{ij} a_{ik} a_{jk}}{\binom{k_i}{2}}$$

- Global clustering:

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

Comments on clustering coefficients I

- Given a network, C and C_{WS} can differ substantially.
- C_{WS} has been used very often for historical reasons (C_{WS} was proposed first).
- C is can be dominated by the contribution of vertices of high degree (which have many adjancent nodes).
- C_{WS} is can be dominated by the contribution of vertices of low degree (which are many in the majority of networks).
- C_{WS} needs taking further decision on C_i when $k_i < 2$ (C is more elegant from a mathematical point of view).

Comments on clustering coefficients II

- Conclusion 0: C and C_{WS} meassure transitivity in different ways (different assumptions/goals).
- Conclusion 1: each measure has its strengths and weaknesses.
- Conclusion 2: explain your methods with precision!

Comments on efficient computation

- Computational challenge: time consuming computation of metrics on large networks.
- Solution: Monte Carlo methods for computing.
- Instead of computing

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

estimate C_{WS} from a mean of C_i over a small fraction of randomly selected vertices.

- High precision exploring a small fraction of nodes (e.g., 5%).

Degree correlations I

What is the dependency between the degrees of vertices at both ends of an edge?

- Assortative mixing (by degree): high degree nodes tend to be connected to high degree nodes, typical of social networks (coauthorship in physics, film actor collaboration,...).
- Disassortative mixing (by degree): high degree nodes tend to be connected to low degree nodes, e.g., neural network (*C. Elegans*), ecological networks (trophic relations).
- No tendency (e.g., Erdös-Rényi graph, Barabási-Albert model).

Degree correlations II

- k_i: degree of the i-th vertex.
- $k'_i = k_i - 1$: remaining degree of the i-th after discounting the edge $i \sim j$.

Correlation

- correlation between k_i and k_j for every edge $i \sim j$.
- correlation between k'_i and k'_j for every edge $i \sim j$.
- metric ρ: $-1 \leq \rho \leq 1$.

Interclass correlation

Theoretical (interclass) correlation:

$$\begin{aligned}\rho(X,Y) &= \frac{COV(X,Y)}{\sigma_X\sigma_Y} \\ &= \frac{E[(X-E[X])(Y-E[Y])]}{\sigma_X\sigma_Y} \\ &= \frac{E[XY]-E[X]E[Y]}{\sigma_X\sigma_Y}\end{aligned}$$

Symmetry: $\rho(X,Y)=\rho(Y,X)$, $\rho_S(X,Y)=\rho_S(Y,X)$.

Empirical correlation:

- Paired mesurements: $(x_1,y_1),...,(x_i,y_i),...,(x_n,y_n)$.
- Sample (interclass) correlation:

$$\rho_s(X,Y)=\frac{\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^n(y_i-\bar{x})^2}}$$

Intraclass correlation

Theoretical intraclass correlation:

$$\rho = \frac{COV_{intra}(X)}{\sigma(X)^2}$$

Empirical correlation:

- Paired measurements: $(x_{1,1}, x_{1,2}),\ldots,(x_{i,1}, x_{i,2}),\ldots,(x_{n,1}, x_{n,2})$

$$\rho_s = \frac{1}{(N-1)\sigma_s^2}\sum_{i=1}^{n}(x_{i,1}-\bar{x})(x_{i,2}-\bar{x})$$

$$\bar{x} = \frac{1}{2N}\sum_{i=1}^{n}(x_{i,1}+x_{i,2})$$

$$\sigma_s^2 = \frac{1}{2(N-1)}\sum_{i=1}^{n}\left[(x_{i,1}-\bar{x})^2+(x_{i,2}-\bar{x})^2\right]$$

Interclass vs intraclass correlation

Interclass correlation:

- Correlation between two variables.

Intraclass correlation:

- Correlation between two different groups (same variable)
- Extent to which members of the same group or class tend to act alike.

Degree correlations III

Intraclass Pearson degree correlation: in an edge $i \sim j$, $X = k_i'$ and $Y = k_j'$ [Newman, 2002].
Three possibilities

- Assortative mixing (by degree): $\rho > 0$, $\rho_s \gg 0$
- Disassortative mixing (by degree): $\rho < 0$, $\rho_s \ll 0$
- No tendency $\rho = 0$, $\rho_s \approx 0$

See Table I of [Newman, 2002] arxiv.org.

General comments on degree correlations I

- *A priori*, a least two ways of measuring degree correlations:
  - $X = k_i$ and $Y = k_j$ (Pearson correlation coefficient)
  - $X = rank(k_i)$ and $Y = rank(k_j)$ (**Spearman** rank correlation)
- rank(k): the smallest k has rank 1, the 2nd smallest k has rank 2 and so on. In case of tie, the degrees in a tie are assigned a mean rank.
- Example:

Sorted degrees	1	3	5	6	6	6	8
The ranks are	1	2	3	$\frac{4+5+6}{3}$	$\frac{4+5+6}{3}$	$\frac{4+5+6}{3}$	7

General comments on degree correlations II

- For historical and sociological reasons, Pearson correlation coefficient has been dominant if not the only approach.
- A test of significance of ρ_S has been missing (potentially problematic for ρ_S close to 0).
- Spearman rank correlation can capture non-linear dependencies.
- Both can fail if the dependency is not monotonic.

General comments on degree correlations II

Some general myths about correlations:

- "ρ_S must be large to be informative" (e.g. $\rho_S > 0.5$).
  - A low value of ρ_S can be significant (very small p-value). Rigorous testing is the key.
  - Low but significant ρ_S can be due to: trends with lots of noise, or clear trends in a narrow domain.
- "No useful information can be extracted from clouds of points". Counterexamples:
  - Vietnam draft (see pp. 248-249 of "Gnuplot in action", by Phillipp K. Janert).
  - Menzerath's law in genomes.

General comments on degree correlations III

The limits of degree correlations

- Degree correlations are global measures.
- The kind of mixing of a vertex might depend on its degree.
- Solution:
  - The mean degree of nearest neighbours of degree k, i.e.

$$\langle k_{nn} \rangle (k)$$

  - An estimate of

$$E[k'|k] = \sum_{k'} k' p(k'|k),$$

    the expected degree k' of 1st neighbours (adjacent nodes) of a node of degree k.

- [Lee et al., 2006]. Statistical properties of sampled networks. Fig. 10 of arxiv.org / Fig. 9 of doi: 10.1103/PhysRevE.73.016102

Lee, S. H., Kim, P.-J., and Jeong, H. (2006).
Statistical properties of sampled networks.
Phys. Rev. E, 73:016102.

Newman, M. E. J. (2002).
Assortative mixing in networks.
Phys. Rev. Lett., 89:208701.

Newman, M. E. J. (2010).
Networks. An introduction.
Oxford University Press, Oxford.

Watts, D. J. and Strogatz, S. H. (1998).
Collective dynamics of 'small-world' networks.
Nature, 393:440–442.