Complex-Network Modelling and Inference

Lecture 8: Graph features (2)

Matthew Roughan
<matthew.roughan@adelaide.edu.au>
https://roughan.info/notes/Network_Modelling/

School of Mathematical Sciences,
University of Adelaide

March 7, 2024

Section 1

Graph features/metrics

Graph Notation

- The network is defined by the *graph*,

$$G(N, E)$$

- We will assume (unless stated) that it is undirected.
- By default label the nodes $\{1, 2, \ldots, n\}$

Graph Features/Metrics

There are two type of metrics/features

- Local (to the nodes)
  - node degree
  - local clustering coefficient
  - centrality (various versions)
  - eccentricity
- Local (to a pair of nodes)
  - (shortest path) distance
- Global (for the whole network)
  - average node degree and degree distribution
  - radius, average distance and diameter
  - global clustering coefficient
  - assortativity/homophily
  - graph spectrum

Section 2

Distance

Distance metrics

A distance metric $d(\cdot, \cdot)$ is function of pairs of elements x, y of a set S to the non-negative real numbers, such that

$$d : S \times S \to [0, \infty),$$

has the properties

1	non-negativity:	$d(x, y) \geq 0$
2	identity:	$d(x, y) = 0 \Leftrightarrow x = y$
3	symmetry:	$d(x, y) = d(y, x)$
4	triangle inequality:	$d(x, z) \leq d(x, y) + d(y, z)$

On a graph, we would like a distance metric on the set of nodes N, *i.e.*, d_{ij} for all $i, j \in N$.

Distances in graphs

- There are many possible distance metrics on a typical graph
- Most are linked to the idea of the "shortest" path
  - provide a distance for each edge
  - distance between two nodes is the sum of the distances of the edges on the shortest path
  - also known as geodesic distance
  - we might say the distance between unconnected nodes is ∞
- *e.g.*,
  - "hop" distance
  - physical links have a distance
  - we will talk in general of "weighted" links, where the weights give distances
- can be generalised (a lot)

Erdős numbers

If you wrote a paper with Erdős, your number is 1. If you wrote a paper with a direct co-author, your number is two, and so on. Essentially it is the graph distance you are from Erdős in a co-authorship graph.

So Erdős number is your "hop" count distance from Erdős is the co-collaborator graph.

http://en.wikipedia.org/wiki/Erdos_number

My Erdős number is 4 (through Charles Pearce, Gavin Brown, and Robert Tijdeman.)

http://www.ams.org/mathscinet/collaborationDistance.html

Metrics associated with distance: average

- Distance is a metric associated with each pair of nodes, so there are $O(|N|^2)$ distances. We usually want to reduce this to a smaller set of measurements
  - most of these assume the graph is connected
- An obvious metric is the *average distance*

$$d_G = \frac{\sum_{i,j \in N} d_{ij}}{n(n-1)}.$$

Metrics associated with distance: eccentricity ...

Definition

The *eccentricity* $\varepsilon(i)$ of a vertex i is the greatest distance between i and any other vertex.

$$\varepsilon(i) = \max_j d_{ij}.$$

- the *radius* of a graph is the minimum eccentricity of any vertex

$$\text{radius}\big(G(N,E)\big) = \min_{i \in N} \varepsilon(i) = \min_i \max_j d_{ij}.$$

- the *diameter* of a graph is the maximum eccentricity of any vertex

$$\text{diameter}\big(G(N,E)\big) = \max_{i \in N} \varepsilon(i) = \max_i \max_j d_{ij}$$

  which is the maximum distance between any pair of nodes.
- a *peripheral vertex* is one whose eccentricity achieves the diameter.
- a *central vertex* is one whose eccentricity achieves the radius

Issues

- Often distance is implicitly a hop count
  - this isn't too interesting to me
  - real networks usually have more meaningful distances
- Distance in directed graphs is not symmetric, so it isn't a formal distance metric
  - *quasi-metrics* are like distance metrics, but give up on symmetry
- In order to calculate distances, we need to calculate shortest paths, which you might not know how to do yet (but we will learn later).

Section 3

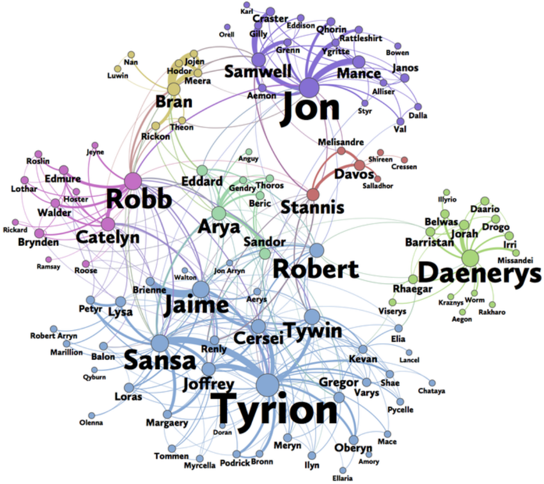
Centrality

Centrality

- We already saw one definition of a "central" node
  - based on distances
  - there are actually multiple competing definitions
- Centrality is associated with importance,
  - *e.g.*, most influential person in a social network or organisation
  - *e.g.*, most important person (or thing) in a movie (the MacGuffin)
  - *e.g.*, a "central" point of failure in a computer network
  - *e.g.*, "super-spreaders" of disease
  - *e.g.*, potential bottlenecks in transport networks

Network of Thrones

Who is the most important character in Game of Thrones?



http://www.npr.org/2016/04/16/474396452/how-math-determines-the-game-of-thrones-protagonist

Metric 4: centrality

- Different measures
  - *Degree centrality*
    - the normalized degree of nodes
    - interpretation — how likely to catch a disease
    - extension to a metric on a graph (maximized by star)
  - *Closeness centrality*
    - reciprocal of mean geodesic distance between x and other nodes

$$c(x) = \frac{1}{\sum_y d(y,x)}$$

  - *Harmonic centrality*
    - mean of reciprocal of geodesic distance between x and other nodes

$$c(x) = \sum_{y \neq x} \frac{1}{d(y,x)}$$

  - *Betweenness centrality*
    - normalized measure of how many shortest-paths a vertex appears on
  - *Eigenvector centrality* $\sim$ Google's PageRank
  - Others: information centrality, cross-clique centrality, percolation centrality, ...

Betweenness centrality

- Quantifies the number of times the node provides "connective tissue" of the graph
- Calculation
    1. Calculate all the shortest paths in the network
    2. Calculate

$$\begin{aligned} \sigma_{st} &= \text{number of shortest paths from } s \text{ to } t \\ \sigma_{st}(x) &= \text{number of shortest paths from } s \text{ to } t \text{ through } x \end{aligned}$$

    3. $$c_B(x) = \frac{1}{K} \sum_{s \neq t \neq x} \frac{\sigma_{st}(x)}{\sigma_{st}}.$$

    where K is total number of possible pairs of vertices not involving x, *e.g.*, in undirected graphs $K = (n-1)(n-2)/2$.

Section 4

Clustering

Clustering

- A key idea is that in many networks we have smaller groups of "clusters"
  - highly connected subnets (*e.g.*, almost cliques)
- For instance, in social networks
  - a friend's friends are more likely to also be my friends
- Clustering metrics assess to which degree a particular network has this property
  - they can be local
  - or global

Global clustering coefficient

- *Clustering coefficient* is a global measure of whether nodes tend to cluster

$$C = 3t_1/t_2,$$

where

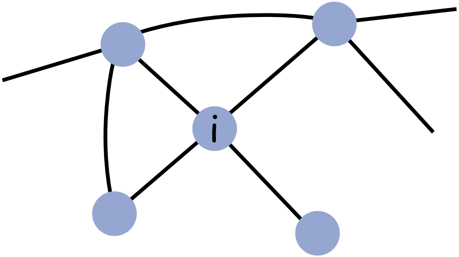
$$t_1 = \text{number of triangles}$$
$$t_2 = \text{number of connected triples or "triplets"}$$

- We take $3t_1$ because each triangle is made up of 3 triplets
- it encodes the idea that in a clustered network it is more likely that a friends' friends are also my friends

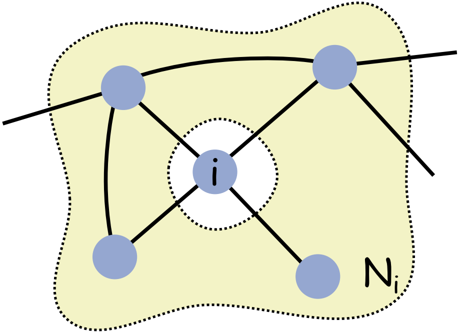
Local clustering coefficient

- Local measure of how close a node and its neigbours are to being a clique
$$c_i = \frac{\left|\{(j,k) \in E \mid j,k \in N_i\}\right|}{k_i(k_i-1)/2},$$
where N_i is the neigbourhood of i, and $k_i = |N_i|$.
- c_i counts the fraction of links in the local neighbourhood, as compared with a clique which has $k_i(k_i-1)/2$
- We can compute a network average clustering co-efficient using
$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} c_i.$$

Local clustering coefficient

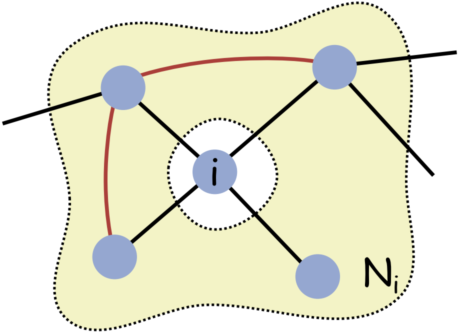


Local clustering coefficient



$$k_i = |N_i| = 4$$

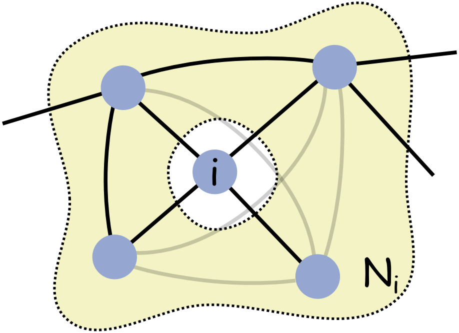
Local clustering coefficient



$$k_i = |N_i| = 4$$

$$\Big|\{(j,k) \in E \mid j,k \in N_i\}\Big| = 2$$

Local clustering coefficient



$$k_i = |N_i| = 4$$

$$\left| \{(j,k) \in E \mid j,k \in N_i\} \right| = 2$$

$$c_i = \frac{1}{3}$$

Section 5

Other metrics

Laplacian and graph spectrum

$$L = D - A$$

- A = adjacency matrix
- D = diagonal matrix of node degrees

Properties

- The eigenvalues of L are sometimes called the **spectrum** of a graph.
- The number of times zero appears in eigenvalues tells you the number of connected components
- resistance distance is related to Moore-Penrose inverse of Laplacian.

Example 1

Human gene regulatory network

Nodes	Genes
Edges	Interactions
$\lvert N \rvert$	21.9 K
$\lvert E \rvert$	12.3 M
$\bar{k}$	1.1 K
Assortativity	0.136
Clustering	0.572

http://networkrepository.com/bio-human-gene1.php

Example 2

IMDB bipartite movie/actor network

Nodes	Movies and actors
Edges	Actor worked in movie
$\|N\|$	896.3 K
$\|E\|$	3.8 M
$\bar{k}$	8
Assortativity	-0.053
Clustering	8.1e-5 [1]

http://networkrepository.com/ca-IMDB.php

[1] Because it is bipartite.

Example 3

Amazon co-purchase network

Nodes	Product
Edges	Co-purchase
$\|N\|$	334.9 K
$\|E\|$	925.9 K
$\bar{k}$	5
Assortativity	-0.059
Clustering	0.205

http://networkrepository.com/com-amazon.php

Yet more metrics

- Metrics specifically for other graphs types
  - reciprocity for digraphs
- Metrics with specific use
  - power-law degree
- Lots of others – for some examples see http://konect.uni-koblenz.de/statistics/

Limitations of metrics

Graphs are complex.

Any small set of numbers will not capture everything important about them.

- *e.g.*, Hamiltonian cycles

Further reading I