Complex-Network Modelling and Inference

Lecture 7: Graph features

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

- When graphs are small, we can draw them, and look at them, and its still hard to assess them, *e.g.*, isomorphism
- When they are large it is impossible to visually assess them
- But we still need ways to analyse them
  - categorise
  - predict
  - assess for unusual characteristics
- There is a growing trend to do so using a common set of numbers called variously
  - statistics
  - metrics
  - features

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

Node degree distributions

Neighbourhood and node degree

Definition

The *neighbourhood* of node i is defined by

$$N_i = \{j \mid (i,j) \in E\},$$

i.e., the set of nodes adjacent to i.

Definition

The *node degree* k_i is the number of neighbours of node i, *i.e.*,

$$k_i = |N_i|.$$

Definition (Alternative definition)

The *node degree* k_i is the number of edges incident to i, *i.e.*,

$$k_i = \Big|\{(i,j)|(i,j) \in E\}\Big|.$$

Global node-degree statistics

- An often used statistic/feature of a graph is its *average node degree*

$$\bar{k} = <k> = \frac{1}{|N|} \sum_{i \in N} k_i = \frac{2|E|}{|N|}.$$

  the last result by the Handshake theorem.
- More generally the *node degree distribution* p_k gives the probability that a node has degree k (or relative frequency)
  - an r-regular graph has

$$p_k = \begin{cases} 1, & \text{for } k = r, \\ 0, & \text{otherwise.} \end{cases}$$

Friendship paradox [Fel91]

Friendship paradox = your friends have more friends than you

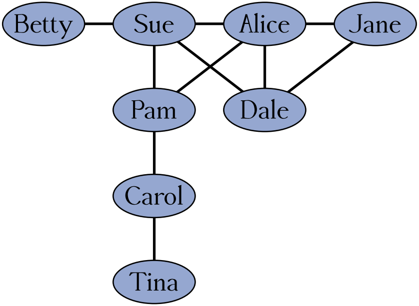
- Actually the theorem is statement about averages
  "On average, your friends will have more friends than you."
- Stated mathematically

$$E[k_i] \leq E[k_{neighbours(i)}].$$

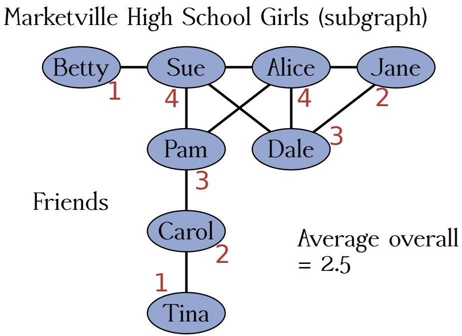
- Some versions are about "most" people, *i.e.*, most people's friends have more friends than them
- Intuition is that sampling the node-degree distribution by looking at friends artificially biases the high-degree nodes because they are friends more often.

Friendship paradox example [Fel91]

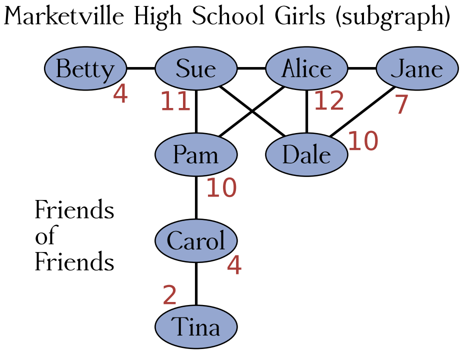
Marketville High School Girls (subgraph)



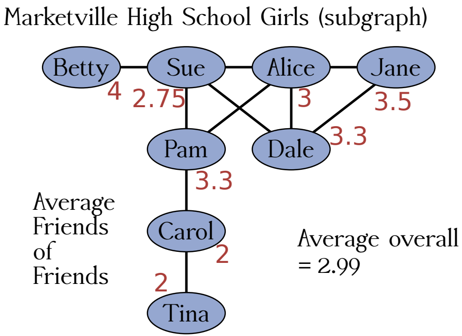
Friendship paradox example [Fel91]



Friendship paradox example [Fel91]



Friendship paradox example [Fel91]



Friendship paradox

Proof.

The average number of friends is

$$\bar{k} = \frac{1}{n} \sum_{i=1}^{n} k_i.$$

When we calculate the number of friends of friends $k_i^{(2)}$, Feld [Fel91] argued that each individual is "a friend k_i times and has k_i friends, so that individual contributes ... a total of k_i^2 friends' friends". Thus the total friends' friends is

$$\sum_{i=1}^{n} k_i^{(2)} = \sum_{i=1}^{n} k_i^2,$$

and we average this over the total number of friends, *i.e.*, $\sum_i k_i$, to get

$$\overline{k^{(2)}} = \frac{\sum_{i=1}^{n} k_i^2}{\sum_{i=1}^{n} k_i}.$$

Friendship paradox

Proof.

Standard result

$$Var(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

Re-arranging we get

$$\frac{\mathbb{E}\left[X^2\right]}{\mathbb{E}[X]} = \mathbb{E}[X] + \frac{Var(X)}{\mathbb{E}[X]}.$$

In our context

$$\overline{k^{(2)}} = \frac{\sum_{i=1}^n k_i^2}{\sum_{i=1}^n k_i} = \mathbb{E}[k_i] + \frac{Var(k_i)}{\mathbb{E}[k_i]},$$

and we know the mean and variance are ≥ 0 so

$$\overline{k^{(2)}} \geq \bar{k}.$$

$\square$

Friendship paradox

Proof (part 2)

- I find the first part a little hand-wavy
- There is a nice little lesson to learn in doing it mathematically

Proof.

We can write the number of friends of i using the adjacency matrix $A = [a_{ij}]$

$$k_i = \sum_j a_{ij} = \sum_j a_{ji}.$$

We can similarly write the number of friends of friends for i using the adjacency matrix by considering that a friend j of i will be reached by a_{ij}, so

$$k_i^{(2)} = \sum_j a_{ij} k_j.$$

$\square$

An aside

The number i's friends' friends can be seen as

$$k_i^{(2)} = \sum_j a_{ij} k_j = \sum_j a_{ij} \sum_k a_{jk} = \sum_k \sum_j a_{ij} a_{jk}$$

Which is just the matrix squared, *i.e.*, if $A^2 = [a_{ij}^{(2)}]$ then we can think of this as

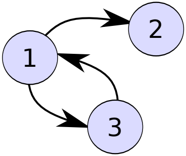
$$a_{ik}^{(2)} = \sum_j a_{ij} a_{jk} = \text{the number of 2 hop paths from } i \text{ to } j,$$

and $k_i^{(2)}$ is the sum over the possible end-points of such paths.

The main point is A^2 contains the number of two-hop paths between each pair of nodes.

An aside: example

Example directed graph



$$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Now

$$A^2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

So there are

- exactly one 2-hop path from 1 to 1: 1-3-1
- exactly one 2-hop path from 3 to 2: 3-1-2
- exactly one 2-hop path from 3 to 3: 3-1-3
- no other 2-hop paths

Friendship paradox

Proof.

$$k_i^{(2)} = \sum_j a_{ij} k_j.$$

So the total number of friends' friends is

$$\sum_i k_i^{(2)} = \sum_i \sum_j a_{ij} k_j.$$

Principle of perversity (of sums) leads us to change the order of summation

$$\sum_i \sum_j a_{ij} k_j = \sum_j \sum_i a_{ij} k_j = \sum_j k_j \sum_i a_{ij} = \sum_j k_j^2.$$

Hence

$$\sum_i k_i^{(2)} = \sum_j k_j^2.$$

$\square$

Section 3

Homophily and Assortativity

Homophily

Birds of a feather, flock together.

- Homophily expresses the idea that many relationships (that we might express in a graph) are more likely between similar entities.
- Many studies have confirmed it in many contexts
  - characteristics: age, gender, class, geography, ...
  - relationships: collaboration, friendship, ...
- In random graphs (see next week) assume the probability of a link depends on similarity of characteristics of the nodes
  - we'll define in terms of e_{ij} = fraction of edges that connect a vertex of type i to one of type j.

Homophily and assortativity

- Assortative mixing, or just *assortativity* expresses homophily between nodes based on their node degree.
- The definition is slightly circular: edges are more common between nodes with similar numbers of edges ...
  - But we can work with that
- Measure using Pearson correlation coefficient of "remaining" degrees at either ends of a random edge.
  - start with the idea of a correlation of nodes on a random link
  - but eliminate the link in question $\Rightarrow$ remaining
  - e_{jk} is the joint probability distribution of the remaining degree at either end of a randomly chosen link

Remaining degree distribution

- The degree distribution of a node reached by a random link

$$q'_k \quad = \quad \frac{k\, p_k}{\sum_j j\, p_j}, \quad k = 1, 2, \ldots$$

- The *remaining degree distribution*, ignores the link we came in on

$$q_k = q'_{k+1} = \frac{(k+1) p_{k+1}}{\sum_j j\, p_j}, \quad k = 0, 1, 2, \ldots$$

- σ_q^2 variance of of distribution q_k

$$\sigma_q^2 = \sum_k k^2 q_k - \left[\sum_k k q_k \right]^2$$

- q_k is the marginal distribution of e_{jk}, *i.e.*,

$$q_k = \sum_j e_{jk}$$

Metric 2: assortativity

- Assortativity

$$r = \frac{\sum_{j,k} jk(e_{jk} - q_j q_k)}{\sigma_q^2}$$

- r is the Pearson correlation coefficient of remaining degrees at either ends of a random edge.

$$-1 \leq r \leq 1$$

- cases
  - r near 1 means high degree nodes often connect to high degree nodes
  - r near -1 means high degree nodes often connect to low degree nodes

Evaluating assortativity

In a real network, we evaluate r by taking

$$\hat{r} = \frac{w\sum_{e\in E} j_e k_e - \left[w\sum_{e\in E}(j_e+k_e)/2\right]^2}{w\sum_{e\in E}(j_e^2+k_e^2)/2 - \left[w\sum_{e\in E}(j_e+k_e)/2\right]^2},$$

where

$$\begin{aligned} w &= |E|^{-1} \\ j_e &= \text{degrees of vertex at one end of the edge } e \\ k_e &= \text{degrees of vertex at other end of the edge } e \end{aligned}$$

Further reading I

Scott L. Feld, *Why your friends have more friends than you do*, American Journal of Sociology **96** (1991), no. 6, 1464–1477.