Complex-Network Modelling and Inference

Lecture 10: Random Graphs: Erdos-Renyi random graphs

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

Random Graphs

Why?

- We often need graphs to use in simulations
  - because we aren't clever enough to do analysis of layers of network protocols on top of a graph
  - *e.g.*, simulations of communications networks
- We need statistical ensembles of graphs to test ideas
  - and there is only 1 real graph
  - *e.g.*, to generate confidence intervals on results
- Random graphs can let us test hypotheses
  - postulate a particular type of random graph as a model
    - sometimes null models or straw men
  - look at its features
- Often want to understand graph behaviour as it gets larger than any examples we have
  - *e.g.*, how will my algorithm work in the future if the network gets much bigger?

The idea at the root

- We start with the idea that there is an ensemble of graphs
  - *e.g.*, $\mathcal{G}_n = \{$all graphs with n nodes$\}$
  - *e.g.*, $\mathcal{G}_{n,k} = \{$all graphs with n nodes and k edges $\}$
  - but these ensembles are usually VERY VERY big
- Then we apply a probability measure to the ensemble, e.g., define

$$P(G), \ \ \forall G \in \mathcal{G}_n$$

But note that

  - $P(G)$ might be too small to calculate
  - $P(G)$ may be too computationally complex to calculate
  - Even if $P(G)$ is easy, we don't want to use it directly
    - *e.g.*, even if we knew $P(G) = const$, we don't want to search through all possible graphs to get "the one"
- So we need a method for constructing graphs that match a given probability distribution, or usually that match some observed features of our graph(s) of interest

Section 2

Gilbert-Erdős-Rényi random graph

Gilbert-Erdős-Rényi random graph [Gil59, ER60]

$$G(n, p)$$

- Take $n = |N|$ nodes
- connect them at random
  - for each pair of nodes flip a (biased) coin
  - if it is heads connect them
- nodes are adjacent with probability p
  - number of edges will be binomial as we have $n(n-1)/2$ iid Bernoulli trials, so

$$\text{prob}(|E| = k) = \binom{n(n-1)/2}{k} p^k (1-p)^{n(n-1)/2-k}.$$

- all graphs with n nodes, and k edges have equal probability

$$P(G|k \text{ edges}) = 1/|\mathcal{G}_{n,k}| = const$$

Gilbert-Erdős-Rényi random graph features

- Average number of links $e = |E|$

$$E[e] = pn(n-1)/2 = p\binom{n}{2}.$$

- Degree distribution is also binomial

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

- critical threshold $np = 1$
  - As p or n increases, the graphs become more and more likely to be connected

Limits of the Binomial distribution: I

Binomial

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}.$$

Take limit as $n \to \infty$, the Binomial distribution approaches a "Normal" distribution $\mathcal{N}(np, np(1-p))$, i.e,

- mean is $\mu = np$
- variance is $\sigma^2 = np(1-p)$
- distribution is Gaussian, i.e.,

$$p(x) \simeq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

Limits of the Binomial distribution: I

Proof: by the Central Limit Theorem which states: take sum of n iid random variables with finite variance

$$S_n = \sum_{i=1}^{n} X_n,$$

then in the limit as $n \to \infty$

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\xrightarrow{d}$ means convergence in distribution. A Binomial distribution is the sum of n iid Bernoulli random variates to the result is immediate.

Limits of the Binomial distribution: II

Binomial

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}.$$

Take limit as $n \to \infty$, such that $np = \lambda$ is kept constant. The Binomial converges to the Poisson distribution:

$$p_k = \frac{\lambda^k e^{-\lambda}}{k!}.$$

- mean is $\lambda = np$
- variance is $\sigma^2 = \lambda = np$

Limits of the Binomial distribution: II

Proof: $np = \lambda$, so $p = \lambda/n \to 0$

$$
\begin{aligned}
p_k &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!} p^k (1-\lambda/n)^{-k} (1-\lambda/n)^n \\
&\simeq \frac{n!}{k!(n-k)!} p^k 1 \exp(-\lambda) \\
&\simeq \frac{n!}{(n-k)!n^k} \frac{\lambda^k}{k!} \exp(-\lambda) \\
&\simeq \frac{\lambda^k}{k!} \exp(-\lambda)
\end{aligned}
$$

Erdős-Rény random graph features

critical piece of information is $np = \lambda$ and how this behaves as n increases

- node degree distribution is approximately Poisson

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \simeq \frac{\lambda^k}{k!} \exp(-\lambda)$$

- average number of links per node is $(n-1)p \simeq \lambda$
  - for $\lambda < 1$, average number of links per node is < 1
  - for $\lambda > 1$, average number of links per node is > 1
- probability degree 0 is $p_0 = \exp(-\lambda)$

Connectivity

- Take case that $n \to \infty$ with $np = \lambda$ fixed.
- Chance that two nodes are adjacent is $p \to 0$.
- What is the chance that they are connected?

Connectivity

What is the chance two nodes are connected by a length 2 path?

$$
\begin{aligned}
& \text{prob}\{i, j \text{ are connected by a length 2 path}\} \\
& \quad = \quad 1 - \text{prob}\{\text{no length 2 path exists from } i \text{ to } j\} \\
& \quad = \quad 1 - \prod_{k \neq i,j} \text{prob}\{\text{path } i - k - j \text{ doesn't exist}\} \\
& \quad = \quad 1 - \prod_{k \neq i,j} (1 - \text{prob}\{\text{path } i - k - j \text{ does exist}\}) \\
& \quad = \quad 1 - (1 - p^2)^{n-2} \\
& \quad = \quad 1 - (1 - (\lambda/n)^2)^{n-2} \\
& \quad \to \quad 0
\end{aligned}
$$

Connectivity

Some crude approximations

$$
\begin{array}{rcl}
\text{prob}\{i,j \text{ are connected by a length 1 path}\} & = & p \\
\text{prob}\{i,j \text{ are connected by a length 2 path}\} & \simeq & (n-2)p^2 \\
\text{prob}\{i,j \text{ are connected by a length 3 path}\} & \simeq & (n-2)(n-3)p^3 \\
 & \vdots & \\
\text{prob}\{i,j \text{ are connected by a length k path}\} & \simeq & n^{k-1}p^k = \lambda^k/n
\end{array}
$$

Sum over all possible path lengths and we get

$$
\text{prob}\{\text{a path exists}\} \simeq (\lambda + \lambda^2 + \cdots + \lambda^{n-1})/n
$$

In the limit as $n \to \infty$ the properties of this depend on whether λ is larger than, or smaller than 1.

Connectivity

$\lambda < 1$

$$
\begin{aligned}
\text{prob}\{\text{a path exists}\} &\simeq (\lambda + \lambda^2 + \cdots + \lambda^{n-1})/n \\
&\simeq \sum_{i=1}^{n-1} \lambda^i / n \\
&\simeq \left(\frac{\lambda^n - \lambda}{\lambda - 1} \right) / n \\
&\to 0
\end{aligned}
$$

Connectivity

$\lambda = 1$

$$
\begin{aligned}
\text{prob}\{\text{a path exists}\} &\simeq (\lambda + \lambda^2 + \cdots + \lambda^{n-1})/n \\
&\simeq \frac{n-1}{n} \\
&\to 1
\end{aligned}
$$

$\lambda > 1$

$$
\begin{aligned}
\text{prob}\{\text{a path exists}\} &\simeq (\lambda + \lambda^2 + \cdots + \lambda^{n-1})/n \\
&> \frac{\lambda^{n-1}}{n} \\
&\to \infty
\end{aligned}
$$

(though obviously a real probability can't go to ∞)

Gilbert-Erdős-Rényi random graph features

critical piece threshold for $np = \lambda$

- $np < 1$: the size of the largest connected component grows as $O(\log n)$
- $np = 1$: the size of the largest connected component grows as $O(n^{2/3})$
- $np > 1$: the largest connected component will have $O(n)$ nodes, and the next largest component will contain no more than $O(\log n)$ nodes.

Gilbert-Erdős-Rényi random graph features

Clustering

- global measure of whether nodes tend to cluster

$$c = 3t_1/t_2,$$

- local measure of how close a node and its neigbours are to being a clique

$$c_i = \frac{|\{(j,k) \in E | j,k \in N_i\}|}{k_i(k_i - 1)/2},$$

where N_i is the neigbourhood of i, and $k_i = |N_i|$.

Global clustering

$$c = 3t_1/t_2,$$

where

$$\begin{aligned} t_1 &= \text{number of triangles} \\ t_2 &= \text{number of connected triples} \end{aligned}$$

- If three nodes are connected, they form a triangle if there is a third link.
- probability of a triangle conditional on the other two links is p.
- in the limit as $n \to \infty$ where $np = const$, the global clustering

$$c \to 0$$

Local clustering

$$c_i = \frac{|\{(j,k) \in E | j,k \in N_i\}|}{k_i(k_i - 1)/2},$$

where N_i is the neigbourhood of i, and $k_i = |N_i|$.

- Conditional on k neighbours, there are $k(k-1)/2$ possible other links.
- Each exists with probability p
- On average $p[k(k-1)/2]$ of these exists
- So as $n \to \infty$

$$\begin{aligned} E[c_i] &= \frac{pk(k-1)/2}{k(k-1)/2} \\ &= p \\ &\to 0 \end{aligned}$$

Gilbert-Erdős-Rényi random graph features

- clustering: Gilbert-Erdős-Rényi RGs don't cluster well
  - intuitively the degree of nodes remains roughly the same
  - more choices for destinations of links
  - so "neighbours" become less densely adjacent

Gilbert-Erdős-Rényi Mark II

- Take $n = |N|$ nodes
- connect them with m edges, randomly assigned
- nodes are adjacent with probability $p = \frac{m}{n(n-1)/2}$
- This is really the Erdős-Rényi graph
- in limit $pn^2 \to \infty$ the two types of Gilbert-Erdős-Rényi graphs have similar properties.

Parameter estimation

- Whenever we have a model we should ask
  - how can I estimate its parameters?
  - what data would I need to do so?
- So parameter estimation (formally part of statistics) should also be part of any modelling toolkit

Parameter estimation for Gilbert-Erdős-Rényi

- The number of edges is a binomial

$$|E| \sim Bin\,(|N|(|N|-1)/2, p)$$

- Sufficient statistics for estimating parameters are $|E|$ and $|N|$
- There are numerous estimators for the parameters of Binomial distributions
  - *e.g.*, MLE (Maximum Likelihood Estimator)

$$\hat{p} = \frac{2|E|}{|N|(|N|-1)}$$

  - Also many ways to compute confidence intervals, etc.

Further reading I

P. Erdős and A. Rényi, *On the evolution of random graphs*, Publications of the Mathematical Institute of the Hungarian Academy of Sciences **5** (1960), 17–61.

E.N. Gilbert, *Random graphs*, Annals of Mathematical Statistics **30** (1959), 1441–1144.