

STOR601 - Research Topics

Structured Data - Networks

Clement Lee

2022-01-21 (Fri)

Welcome

- ▶ Outline of this hour
 - ▶ Introduction to statistical network modelling
 - ▶ Stochastic block models
 - ▶ Modelling degree distribution

Introduction to statistical network modelling

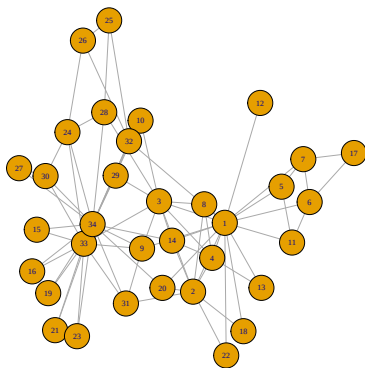
From traditional data

- ▶ We know quite a lot about:
 - ▶ data that are independent & identically distributed (IID)
 - ▶ temporal and/or spatial data
- ▶ Network data doesn't belong to either
 - ▶ Of interest is the relationship **between** nodes in the network
 - ▶ The structure **is** the data

Examples

- ▶ Social networks
 - ▶ Facebook friends, Twitter followers, ...
- ▶ Enron emails
 - ▶ Over 0.6 million emails by 158 employees before collapse in 2001
- ▶ Co-authorship & citation networks (Ji and Jin, 2016)
 - ▶ Academics write papers with other academics
 - ▶ Papers cite other papers
- ▶ Brain networks
 - ▶ Connectome project

The Karate Club (Zachary, 1977) DOI: 10.1086/jar.33.4.3629752



Terminology

- ▶ Nodes & edges
- ▶ Undirected network
 - ▶ Degree of a node: total number of edges
- ▶ Directed network
 - ▶ In-degree: number of edges pointing towards the node
 - ▶ Out-degree: number of edges pointing from the node

Why bother?

- ▶ **Describe** the network in a concise way
 - ▶ $\sim$ probability distributions describing IID data
- ▶ **Examine** how networks grow
 - ▶ Come up with simple rules of how nodes join & connect with others
 - ▶ Allow simulation of networks easily
- ▶ **Discover / understand** underlying structure
 - ▶ Cluster the nodes in a meaningful way
 - ▶ Project them to a 2-D or 3-D space

How?

1. Assume a probabilistic / statistical model
 - ▶ Rules or equations describe how the network comes into being
 - ▶ Parameters characterise the model
2. Fit the model to the data
 - ▶ The usual steps of modelling & inference
 - ▶ Estimate the parameters
 - ▶ Determine if the model fits well
3. Interpret and/or predict
 - ▶ Less well-defined
 - ▶ How it will continue to grow?

Erdos-Renyi Model

- ▶ Some call it Bernoulli random graph
- ▶ Special case of most network models
- ▶ Simple rules:
 - ▶ There are n nodes
 - ▶ Every pair of nodes has the same probability p of connecting / having an edge
 - ▶ And the occurrence of all these edges are independent of each other
- ▶ The degree (number of edges) of a node is a random variable
 - ▶ Follows Binomial($n - 1, p$) distribution
 - ▶ Same for all nodes

What other models?

- ▶ **Describe** the network in a concise way
 - ▶ Exponential random graph models (ERGMs)
- ▶ **Examine** how networks grow
 - ▶ *Mechanistic or generative models*
 - ▶ Preferential attachment model
 - ▶ Small-world model
- ▶ **Discover / understand** underlying structure
 - ▶ Stochastic block models
 - ▶ Latent space models
 - ▶ Latent feature models
 - ▶ Hypergraph models
 - ▶ Graphons

Networks or (random) graphs?

- ▶ Network is the actual data / what we visualise
- ▶ Graph is the mathematical object
 - ▶ How to write it down? One way is the **adjacency matrix**
 - ▶ Let's use the Karate Club as an example

Challenges

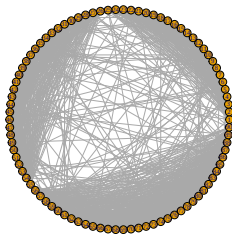
- ▶ Sampling & design
 - ▶ When to stop sampling? How to *cut* the network?
- ▶ Description & visualisation
 - ▶ Lack of predefined geometry
- ▶ Modelling & inference
 - ▶ Same data, different models - which one is the best?
- ▶ Computational cost for big data
 - ▶ Not very scalable

Networks as models?

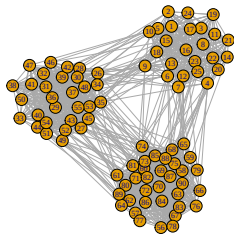
- ▶ So far the networks are the **data**
 - ▶ ... and are modelled by statistical models
- ▶ Different to probabilistic models being represented as a network / graph
 - ▶ Graphical model
 - ▶ Bayesian network
- ▶ Alex Gibberd will talk about graphical models

Stochastic block models

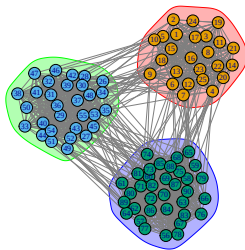
Example: 90 nodes, 1192 edges



Visually better



Colouring the groups



Observations

- ▶ Nodes are tightly connected to each other **within** each group
 - ▶ Less connected **between** groups
- ▶ More importantly, the connection pattern is **similar** for all nodes within the same group
 - ▶ Approx. same number of edges with nodes in the same group
 - ▶ Approx. same number of edges with nodes in another group
- ▶ I simulated the data & made the communities visible
 - ▶ Less clear-cut for real-life data

Motivation

- ▶ Given: a network
 - ▶ Undirected or directed
 - ▶ There could be node/edge attributes
 - ▶ **Not** considered as classified labels though
- ▶ To discover:
 - ▶ Groups in which nodes are tightly connected (community detection); **and/or**
 - ▶ Groups in which the connectivity pattern is similar
- ▶ Node clustering / unsupervised learning

Community detection **algorithms**

- ▶ Related to stochastic block models
- ▶ Heuristic algorithms
- ▶ Quick, based on simple rules
- ▶ Comprehensive review by Fortunato (2010, Physics Reports)
- ▶ Let's look at an example

Label propagation algorithm

- ▶ Raghavan, Albert and Kumara (2007, *Physical Review E*)
- ▶ Initially each node is randomly assigned to one of the K groups
- ▶ Each node takes turn to join the group to which the maximum number of neighbours belong
- ▶ The iterative process continues until no node changes its group membership anymore
- ▶ Cross reference: k -nearest neighbours algorithm

Issues of community detection algorithms

- ▶ Different initial configurations may lead to different results
- ▶ Different algorithms give different results
 - ▶ Which one is better / best?
- ▶ Pre-specifying the number of groups beforehand

Stochastic block models (SBMs)

- ▶ Proper statistical models
 - ▶ Theoretically one set of results, once the model is specified
- ▶ Idea: To put nodes that have **similar** properties together
 - ▶ The concept of **stochastic equivalence**

Stochastic equivalence

- ▶ Given nodes p and q in the same group
- ▶ $\Pr(\text{edge from } p \text{ to } r) = \Pr(\text{edge from } q \text{ to } r)$
 - ▶ No matter r is in the same group with p & q or not
 - ▶ The two probabilities are independent, **given the group memberships**
- ▶ This is the one I used for the example

	Group 1	Group 2	Group 3
Group 1	0.8	0.05	0.05
Group 2	0.05	0.8	0.05
Group 3	0.05	0.05	0.8

SBMs do not necessarily achieve community detection

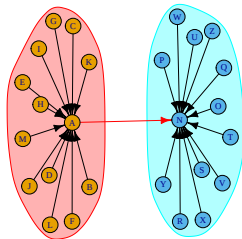
- ▶ The model equation says nothing about the probabilities' magnitude
- ▶ High number of connections between groups but not within groups

	Group 1	Group 2	Group 3
Group 1	0.03	0.06	0.80
Group 2	0.90	0.05	0.01
Group 3	0.02	0.70	0.04

Fitting to real networks

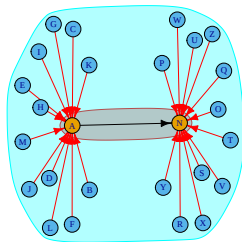
- ▶ SBMs *usually* achieve community detection
 - ▶ Able to cluster tightly connected nodes
- ▶ This means the following are assumed:
 - ▶ Stochastic equivalence for tightly connected nodes
 - ▶ Degrees of nodes within each group follow the same distribution
 - ▶ Quite strong assumptions
- ▶ Not quite realistic for actual data
 - ▶ Degree heterogeneity
 - ▶ Lack of fit

Illustration



- Community detection achieved but poor fit

Illustration (cont'd)



- Nodes stochastic equivalent but not tightly connected

Degree-corrected SBM

- ▶ Karrer and Newman (2011, *Physical Review E*)
- ▶ An extra parameter / latent variable for each node
 - ▶ To correct for the heterogeneity in their degrees
 - ▶ Nodes within the same group “stochastically equivalent” after correcting
- ▶ Community detection + improved fit

Not the universal fix though

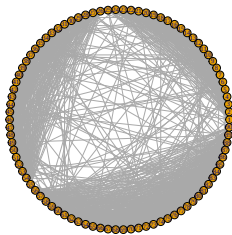
- ▶ In some cases, selecting between the original and degree-corrected versions is not clear cut
- ▶ Each version reveals a different underlying structure about the same network
- ▶ Instead of choosing between the two versions, they could be combined in a weighted / mixture model

Inference

- ▶ What data do we have?
 - ▶ The network, represented by e.g. the adjacency matrix
- ▶ What quantities do we have to infer?
 - ▶ The group memberships of the nodes
 - ▶ The block matrix

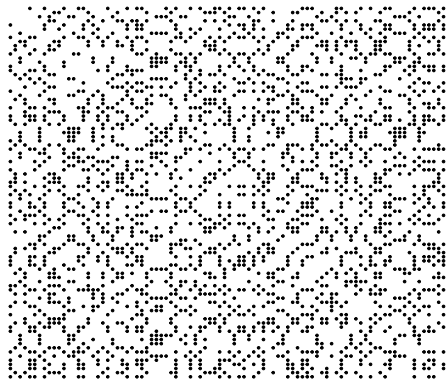
	Group 1	Group 2	Group 3
Group 1	??	??	??
Group 2	??	??	??
Group 3	??	??	??

Using the first example



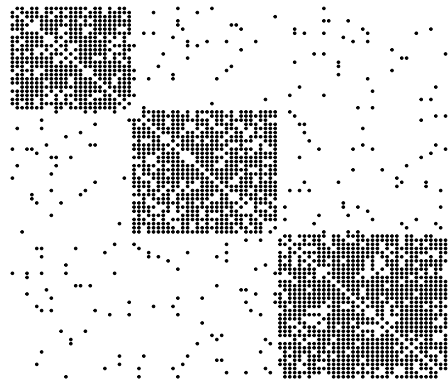
Rearranging the adjacency matrix

► From ...



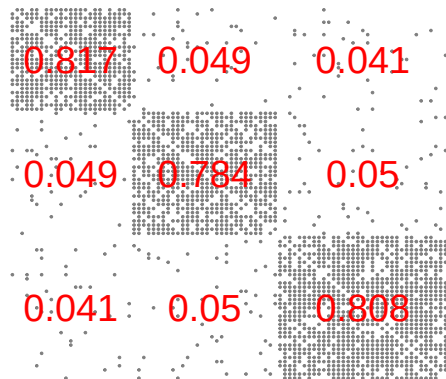
Inference achieved

► ... to ...



Adjacency matrix $\rightarrow$ block matrix

- Densities in each block = point estimates of the parameters



Inference issues

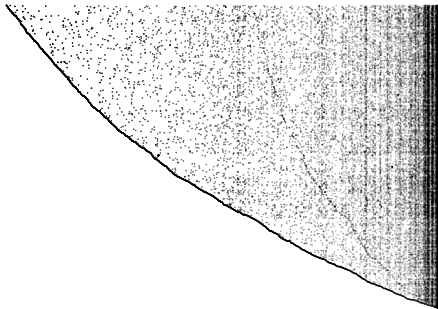
- ▶ Computational cost of naive inference algorithm usually $O(n^2)$
 - ▶ n = number of nodes
- ▶ Number of groups K
 - ▶ Infer it as a parameter?
 - ▶ Select it using criteria?
- ▶ Label switching issue
 - ▶ Common in clustering
- ▶ Bayesian approach by Peixoto (2020, *ArXiv:1705.10225*)
 - ▶ Biggest development since the degree-corrected SBM
 - ▶ Seems to solve most modelling & inference issues

Transition: my work

- ▶ Lee and Wilkinson (2019, *Applied Network Science*)
 - ▶ Review of stochastic block models
 - ▶ All references before (& more)
- ▶ Networks that are **directed acyclic graphs** (DAGs)
 - ▶ Citation network of all the papers in a conference
 - ▶ Dependency network of R packages

Modifying SBM for DAGs

- Adjacency matrix can be upper triangular upon rearranging



Investigating the degree heterogeneity

- ▶ Instead of considering it in the context of SBMs
- ▶ How about we look at the degree distribution directly?

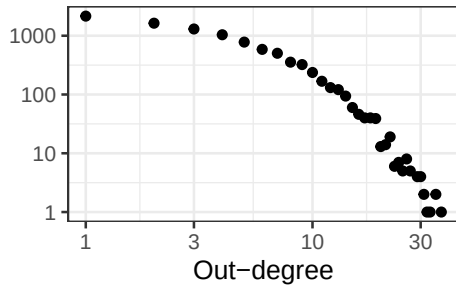
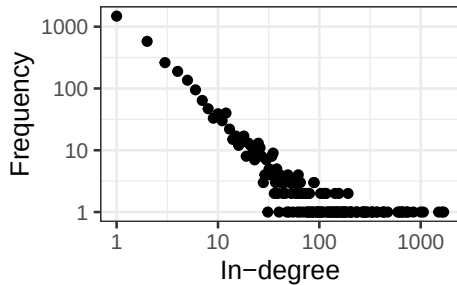
Modelling degree distribution

What does degree heterogeneity mean?

- ▶ Nodes not following the same degree distribution
- ▶ There may be underlying reasons for such heterogeneity
 - ▶ (Unknown) node covariates
 - ▶ But they are not the focus here
- ▶ Simply fit a univariate distribution to each kind of degrees:
 - ▶ In-degree
 - ▶ Out-degree
 - ▶ Total degree

Degrees of R packages

- In-degree seems to follow the **power law**



The power law

- ▶ In probability terms:

$$p(x) \propto x^{-\alpha}, \quad x > u$$
$$\log p(x) = -\alpha \log x + c$$

- ▶ Straight line on the log-log scale
 - ▶ $\alpha = -1 \times \text{slope}$
- ▶ Related concepts:
 - ▶ “The rich gets richer”
 - ▶ Preferential attachment
 - ▶ 80-20 rule / Pareto principle
 - ▶ Scale-free networks

Preferential attachment model

- ▶ Barabasi and Albert (1999, *Science*)
- ▶ Growing the network by simple rules
 - ▶ A new node comes in and connect to existing nodes
 - ▶ Probability of connecting to the new node $\propto$ current (in-)degree
- ▶ Degree distribution follows the power law with $2 < \alpha < 3$
 - ▶ Clauset et al. (2009, *SIAM Review*)
- ▶ These two papers are themselves embodiment of preferential attachment

Criticisms of the power law

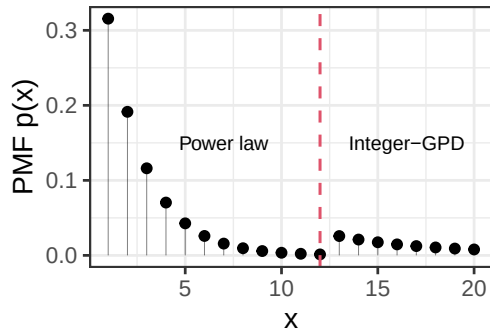
- ▶ Not adequate to describe the degree distribution
 - ▶ Usually poorer fit in the right tail, where it is more of interest
- ▶ Alternative distributions e.g. log-normal better / sufficient
 - ▶ Stumpf and Porter (2012, *Science*)
 - ▶ Broido and Clauset (2019, *Nature Communications*)

Connecting to extremes

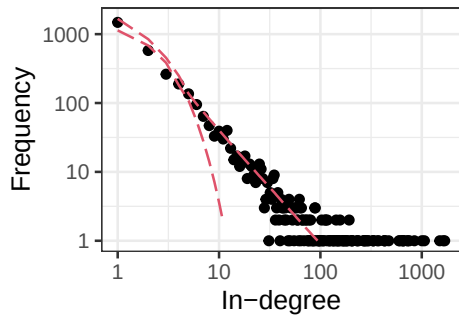
- ▶ Continuous power law = Pareto distribution
- ▶ Pareto distribution → Generalised Pareto distribution (GPD)
 - ▶ Popular in extreme value theory
 - ▶ Can model the tail better / flexibly
- ▶ GPD → Integer-GPD
 - ▶ Discretised
 - ▶ Rohrbeck et al. (2018, *AoAS*)
- ▶ Why don't we use this then?

Sidetrack: extreme value theory

- ▶ Mixture distribution
- ▶ Power law for the left “tail”
 - ▶ $\text{data} \leq u$
- ▶ GPD for the right tail
 - ▶ $\text{data} > u$
 - ▶ More flexible
- ▶ u is also inferred



Good fit



Next steps

- ▶ Can we modify the preferential attachment model to lead to such in-degree distribution?
- ▶ How about modifying an SBM? Or a combination of both?
- ▶ Work in progress: Lee and Eastoe (2020, *ArXiv:2008.03073*)

Interested and want to explore?

- ▶ Literature in the slides
- ▶ R packages
 - ▶ igraph: <https://igraph.org/r/>
 - ▶ visNetwork: <https://cran.r-project.org/package=visNetwork>
 - ▶ ggraph: <https://exts.ggplot2.tidyverse.org/ggraph.html>
- ▶ Other languages
 - ▶ igraph with Python / C: <https://igraph.org/>
 - ▶ vis.js with JavaScript: <https://visjs.org/>
 - ▶ D3.js with JavaScript: <https://observablehq.com/@d3/gallery>

Thank you

► Questions?