Network Modeling

Gonzalo Mateos

Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.ece.rochester.edu/~gmateosb/

Acknowledgments: E. Kolaczyk, F. La Rocca and A. Ribeiro

Facultad de Ingeniería, UdelaR
Montevideo, Uruguay
February 4, 2021

What is this lecture about?

- Statistial graph models are used for a variety of reasons:

1) Mechanisms explaining properties observed on real-world networks
   Ex: small-world effects, power-law degree distributions

2) Testing for 'significance' of a characteristic $\eta(G)$ in a network graph
   Ex: is the observed average degree unusual or anomalous?

3) Assessment of factors potentially predictive of relational ties
   Ex: are there reciprocity or transitivity effects in play?

- Focus today on construction and use of models for network data

Modeling network graphs

- **Def:** A model for a network graph is a collection

$$\{P_\theta(G), G \in \mathcal{G} : \theta \in \Theta\}$$

  - $\mathcal{G}$ is an ensemble of possible graphs
  - $P_\theta(\cdot)$ is a probability distribution on $\mathcal{G}$ (often write $P(\cdot)$)
  - Parameters θ ranging over values in parameter space Θ

- Richness of models derives from how we specify $P_\theta(\cdot)$

  $\Rightarrow$ Methods range from the simple to the complex

Model specification

1) Let $\mathrm{P}(\cdot)$ be uniform on $\mathcal{G}$, add structural constraints to $\mathcal{G}$
   Ex: Erdös-Renyi random graphs, generalized random graph models
2) Induce $\mathrm{P}(\cdot)$ via application of simple generative mechanisms
   Ex: small world, preferential attachment, copying models
3) Model structural features and their effect on G's topology
   Ex: exponential random graph models
4) Model propensity towards establishing links via latent variables
   Ex: stochastic block models, graphons, random dot product graphs

- Computational cost of associated inference algorithms relevant

Roadmap

Classical random graph models

- Assign equal probability on all undirected graphs of given order and size
  - Specify collection $\mathcal{G}_{N_v,N_e}$ of graphs $G(\mathcal{V},\mathcal{E})$ with $|\mathcal{V}| = N_v$, $|\mathcal{E}| = N_e$
  - Assign $\mathrm{P}(G) = \binom{N}{N_e}^{-1}$ to each $G \in \mathcal{G}_{N_v,N_e}$, where $N = |\mathcal{V}^{(2)}| = \binom{N_v}{2}$
- Most common variant is the Erdös-Renyi random graph model $G_{n,p}$

  $\Rightarrow$ Undirected graph on $N_v = n$ vertices

  $\Rightarrow$ Edge (u, v) present w.p. p, independent of other edges
- Simulation: simply draw $N = \binom{N_v}{2} \approx N_v^2/2$ i.i.d. Ber(p) RVs
  - Inefficient when $p \sim N_v^{-1} \Rightarrow$ sparse graph, most draws are 0
  - Skip non-edges drawing Geo(p) i.i.d. RVs, runs in $O(N_v + N_e)$ time

Properties of $G_{n,p}$

- $G_{n,p}$ is well-studied and tractable. Noteworthy properties:

P1) Degree distribution $\mathrm{P}(d)$ is binomial with parameters $(n-1, p)$
- Large graphs have concentrated $\mathrm{P}(d)$ with exponentially-decaying tails

P2) Phase transition on the emergence of a giant component
- If $np > 1$, $G_{n,p}$ has a giant component of size $O(n)$ w.h.p.
- If $np < 1$, $G_{n,p}$ has components of size only $O(\log n)$ w.h.p.

P3) Small clustering coefficient $O(n^{-1})$ and short diameter $O(\log n)$ w.h.p.

Generalized random graph models

- Recipe for generalization of Erdös-Renyi models
  - $\Rightarrow$ Specify $\mathcal{G}$ of fixed order N_v, possessing a desired characteristic
  - $\Rightarrow$ Assign equal probability to each graph $G \in \mathcal{G}$
- Configuration model: fixed degree sequence $\{d_{(1)}, \ldots, d_{(N_v)}\}$
  - Size fixed under this model, since $N_e = \bar{d}N_v/2 \Rightarrow \mathcal{G} \subset \mathcal{G}_{N_v,N_e}$
  - Equivalent to specifying model via conditional distribution on $\mathcal{G}_{N_v,N_e}$
- Configuration models useful as reference, i.e., 'null' models

  Ex: compare observed G with $G' \in \mathcal{G}$ having power law $\mathrm{P}(d)$

  Ex: expected group-wise edge counts in modularity measure

Results on the configuration model

P1) Phase transition on the emergence of a giant component

- Condition depends on first two moments of given $\mathrm{P}(d)$
- Giant component has size $O(N_v)$ as in $G_{N_v,p}$

M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random Struct. and Alg.*, vol. 6, pp. 161-180, 1995

P2) Clustering coefficient vanishes slower than in $G_{N_v,p}$

M. Newman et al, "Random graphs with arbitrary degree distrbutions and their applications", *Physical Rev. E*, vol. 64, p. 26,118, 2001

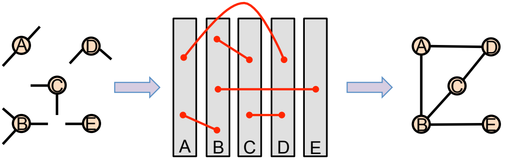
P3) Special case of given power-law degree distribution $\mathrm{P}(d) \sim Cd^{-\alpha}$

- For $\alpha \in (2, 3)$, short diameter $O(\log N_v)$ as in $G_{N_v,p}$

F. Chung and L. Lu, "The average distances in random graphs with given expected degrees," *PNAS*, vol. 99, pp. 15,879-15,882, 2002

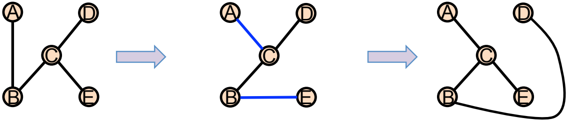
Simulating generalized random graphs

- Matching algorithm



Given: nodes with spokes Randomly match mini-nodes Sample graph

- Switching algorithm



Initialize

Randomly switch a pair of edges
Repeat ~100N_e times

Sample graph

Task 1: Model-based estimation in network graphs

- Consider a sample G^* of a population graph $G(\mathcal{V}, \mathcal{E})$

  $\Rightarrow$ Suppose a given characteristic $\eta(G)$ is of interest

  $\Rightarrow$ **Q:** Useful estimate $\hat{\eta} = \hat{\eta}(G^*)$ of $\eta(G)$?

- Statistical inference in sampling theory via design-based methods

  $\Rightarrow$ Only source of randomness is due to the sampling design

- Augment this perspective to include a model-based component
  - Assume G drawn uniformly from the collection $\mathcal{G}$, prior to sampling

- Inference on $\eta(G)$ should incorporate both randomness due to

  $\Rightarrow$ Selection of G from $\mathcal{G}$ and sampling G^* from G

Directly modeling $\eta(G)$

- So far considered modeling G for model-based estimation of $\eta(G)$

  $\Rightarrow$ Alternatively, one may specify a model for $\eta(G)$ directly

Example

- Estimate the power-law exponent $\eta(G) = \alpha$ from degree counts
- A power law implies the linear model $\log \mathrm{P}(d) = C - \alpha \log d + \epsilon$

  $\Rightarrow$ Could use a model-based estimator such as least squares

- Better form the MLE for the model $f(d;\alpha) = \frac{\alpha-1}{d_{\min}} \left(\frac{d}{d_{\min}}\right)^{-\alpha}$

$$\text{Hill estimator} \quad \Rightarrow \hat{\alpha} = 1 + \left[\frac{1}{N_v}\sum_{i=1}^{N_v} \log\left(\frac{d_i}{d_{\min}}\right)\right]^{-1}$$

Task 2: Assessing significance in network graphs

- Consider a graph G^{obs} derived from observations
- **Q:** Is a structural characteristic $\eta(G^{obs})$ significant, i.e., unusual?
  - $\Rightarrow$ Assessing significance requires a frame of reference, or null model
  - $\Rightarrow$ Random graph models often used in setting up such comparisons
- Define collection $\mathcal{G}$, and compare $\eta(G^{obs})$ with values $\{\eta(G) : G \in \mathcal{G}\}$
  - $\Rightarrow$ Formally, construct the reference distribution

$$\mathrm{P}_{\eta,\mathcal{G}}(t) = \frac{|\{G \in \mathcal{G} : \eta(G) \leq t\}|}{|\mathcal{G}|}$$

- If $\eta(G^{obs})$ found to be sufficiently unlikely under $\mathrm{P}_{\eta,\mathcal{G}}(t)$
  - $\Rightarrow$ Evidence against the null H_0: G^{obs} is a uniform draw from $\mathcal{G}$

Example: Zachary's karate club

- Zachary's karate club has clustering coefficient $\mathrm{cl}(G^{obs}) = 0.2257$

  $\Rightarrow$ Random graph models to assess whether the value is unusual

- Construct two 'comparable' abstract frames of reference
  1) Collection $\mathcal{G}_1$ of random graphs with same $N_v = 34$ and $N_e = 78$
  2) Add the constraint that $\mathcal{G}_2$ has the same degree distribution as G^{obs}

- $|\mathcal{G}_1| \approx 8.4 \times 10^{96}$ and $|\mathcal{G}_2|$ much smaller, but still large

  $\Rightarrow$ Enumerating $\mathcal{G}_1$ intractable to obtain $\mathrm{P}_{\eta,\mathcal{G}_1}(t)$ exactly

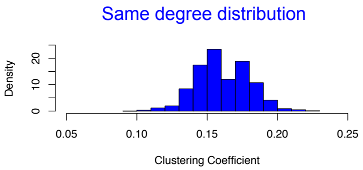
- Instead use simulations to approximate both distributions

  $\Rightarrow$ Draw 10,000 uniform samples G from each $\mathcal{G}_1$ and $\mathcal{G}_2$

  $\Rightarrow$ Calculate $\eta(G) = \mathrm{cl}(G)$ for each sample, plot histograms

Example: Zachary's karate club (cont.)

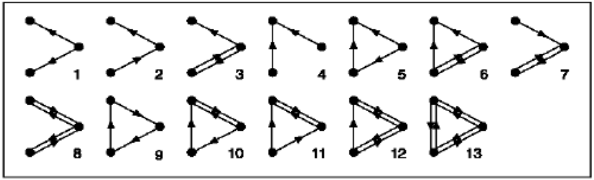
- Plot histograms to approximate the distributions



- Unlikely to see a value $\text{cl}(G^{obs}) = 0.2257$ under both graph models
  Ex: only 3 out of 10,000 samples from $\mathcal{G}_1$ had $\text{cl}(G) > 0.2257$
- Strong evidence to reject G^{obs} obtained as sample from $\mathcal{G}_1$ or $\mathcal{G}_2$

Task 3: Detecting network motifs

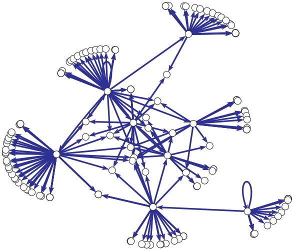
- Related use of random graph models is for detecting network motifs
  - ⇒ Find the simple 'building blocks' of a large complex network
- **Def:** Network motifs are small subgraphs occurring far more frequently in a given network than in comparable random graphs
- Ex: there are $L_3 = 13$ different connected 3-vertex subdigraphs

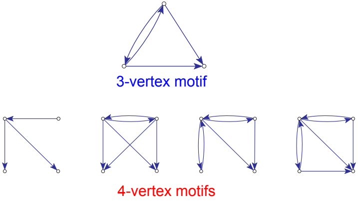


- Let N_i be the count in G of the i-th type k-vertex subgraph, $i = 1, \ldots, L_k$
  - ⇒ Each value N_i can be compared to a suitable reference $\mathbb{P}_{N_i,\mathcal{G}}$
  - ⇒ Subgraphs for which N_i is extreme are declared as network motifs

Example: AIDS blog network

- AIDS blog network G^{obs} with $N_v = 146$ bloggers and $N_e = 183$ links
  - $\Rightarrow$ Examined evidence for motifs of size $k = 3$ and 4 vertices



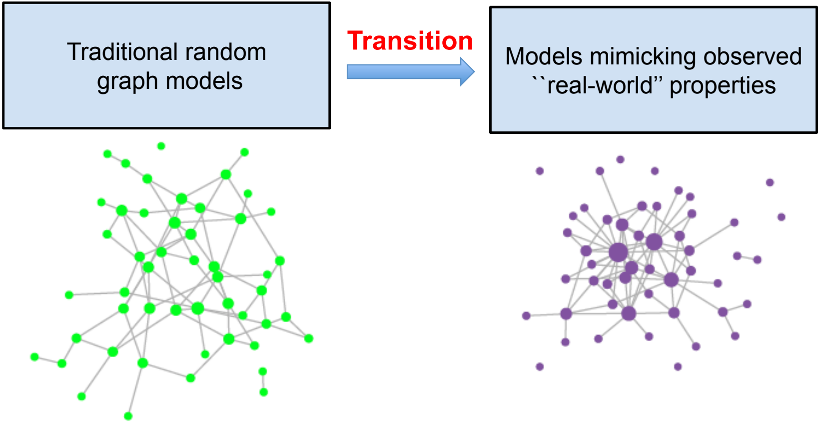


- Simulated 10,000 digraphs using a switching algorithm
  - $\Rightarrow$ Fixed in- and out-degree sequences, mutual edges as in G^{obs}
  - $\Rightarrow$ Constructed approximate reference distributions $\mathrm{P}_{N_i,\mathcal{G}}(t)$
- Ex: two bloggers with a mutual edge and a common 'authority'

Roadmap

Models for real-world networks

- Noteworthy innovation in 'modern' graph modeling



A "small" world?

- Six degrees of separation popularized by a play [Guare'90]
  - ⇒ Short paths between us and everyone else on the planet
  - ⇒ Term relatively new, the concept has a long history
- Traced back to F. Karinthy in the 1920s
  - ⇒ 'Shrinking' modern world due to increased human connectedness
  - ⇒ Challenge: find someone whose distance from you is > 5
  - ⇒ Inspired by G. Marconi's Nobel prize speech in 1909
- First mathematical treatment [Kochen-Pool'50]
  - ⇒ Formally modeled the mechanics of social networks
  - ⇒ But left 'degrees of separation' question unanswered
- Chain of events led to a groundbreaking experiment [Milgram'67]

Milgram's experiment

- **Q1:** What is the typical geodesic distance between two people?

  ⇒ Experiment on the global friendship (social) network

  ⇒ Cannot measure in full, so need to probe explicitly

- S. Milgram's ingenious small-world experiment in 1967
  - 296 letters sent to people in Wichita, KS and Omaha, NE
  - Letters indicated a (unique) contact person in Boston, MA
  - Asked them to forward the letter to the contact, following rules

- **Def:** friend is someone known on a first-name basis

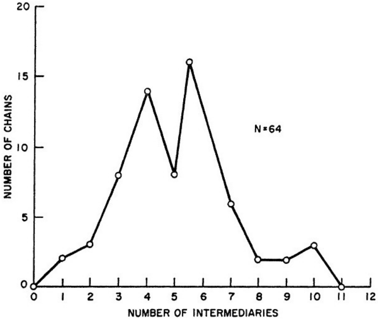
  Rule 1: If contact is a friend then send her the letter; else

  Rule 2: Relay to friend most-likely to be a contact's friend

- **Q2:** How many letters arrived? How long did they take?

Milgram's experimental results

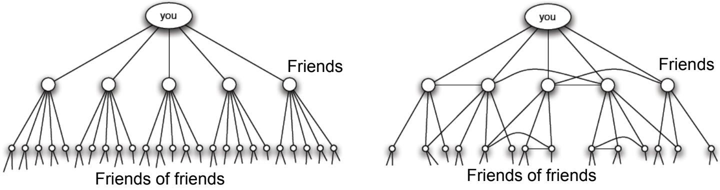
- 64 of 296 letter reached the destination, average path length $\bar{\ell} = 6.2$
  - $\Rightarrow$ Inspiring Guare's '6 degrees of separation'
- Conclusion: short paths connect arbitrary pairs of people



S. Milgram, "The small-world problem," *Psychology Today*, vol. 2, pp. 60-67, 1967

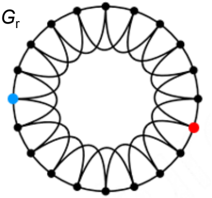
Moment to reflect

- Milgram demonstrated that short paths are in abundance
- **Q:** Is the small-world theory reasonable? Sure, e.g., assumes:
  - We have 100 friends, each of them has 100 other friends, . . .
  - After 5 degrees we get 10^{10} friends $>$ twice the Earth's population

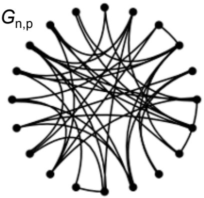


- Not a realistic model of social networks exhibiting:
  - ⇒ Homophily [Lazarzfeld'54]
  - ⇒ Triadic closure [Rapoport'53]
- **Q:** How can networks be highly-structured locally **and** globally small?

Structure and randomness as extremes



High clustering and diameter



Low clustering and diameter

- One-dimensional regular lattice G_r on N_v vertices
  - Each node is connected to its $2r$ closest neighbors (r to each side)

  Structure yields high clustering and high diameter

$$\text{cl}(G_r) = \frac{3r-3}{4r-2} \text{ and } \text{diam}(G_r) = \frac{N_v}{2r}$$

- Other extreme is a $G_{N_v,p}$ random graph with $p = O(N_v^{-1})$

  Randomness yields low clustering and low diameter

$$\text{cl}(G_{N_v,p}) = O(N_v^{-1}) \text{ and } \text{diam}(G_{N_v,p}) = O(\log N_v)$$

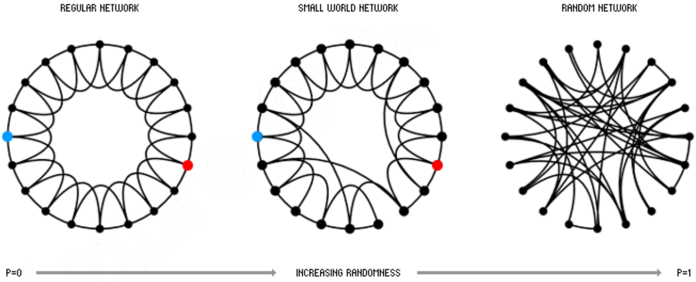
The Watts-Strogatz model

- Small-world model: blend of structure with little randomness

  S1: Start with regular lattice that has desired clustering

  S2: Introduce randomness to generate shortcuts in the graph

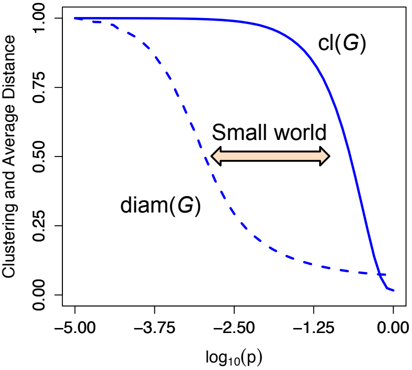
  ⇒ Each edge is randomly rewired with (small) probability p



- Rewiring interpolates between the regular and random extremes

Numerical results

- Simulate Watts-Strogatz model with $N_v = 1,000$ and $r = 6$
  - Rewiring probability p varied from 0 (lattice G_r) to 1 (random $G_{N_v,p}$)
  - Normalized $\mathrm{cl}(G)$ and $\mathrm{diam}(G)$ to maximum values ($p = 0$)



- Broad range of $p \in [10^{-3}, 10^{-1}]$ yields small $\mathrm{diam}(G)$ and high $\mathrm{cl}(G)$

Closing remarks

- Structural properties of Watts-Strogatz model [Barrat-Weigt'00]

  P1: Large N_v analysis of clustering coefficient

$$\text{cl}(G) \approx \frac{3r-3}{4r-2}(1-p^3) = \text{cl}(G_r)(1-p^3)$$

  P2: Degree distribution concentrated around $2r$

- Small-world graph models of interest across disciplines
- Particularly relevant to 'communication' in a broad sense
  - $\Rightarrow$ Spread of news, gossip, rumors
  - $\Rightarrow$ Spread of natural diseases and epidemics
  - $\Rightarrow$ Search of content in peer-to-peer networks

Roadmap

Random graph models

Small-world models

Network-growth models

Exponential random graph models

Latent network models

Random dot product graphs

Time-evolving networks

- Many networks grow or otherwise evolve in time

  Ex: Web, scientific citations, Twitter, genome ...

- General approach to model construction mimicking network growth
  - Specify simple mechanisms for network dynamics
  - Study emergent structural characteristics as time $t \to \infty$

- **Q:** Do these properties match observed ones in real-world networks?

- Two fundamental and popular classes of growth processes

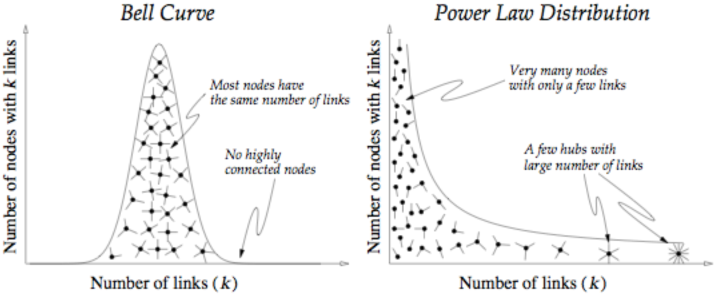
  $\Rightarrow$ Preferential attachment models

  $\Rightarrow$ Copying models

- Tenable mechanisms for popularity and gene duplication, respectively

Popularity as a network phenomenon

- Popularity is a phenomenon characterized by extreme imbalances
  - How can we quantify these imbalances? Why do they arise?



- Basic models of network behavior can be very insightful

  $\Rightarrow$ Result of coupled decisions, correlated behavior in a population

Preferential attachment model

- Simple model for the creation of e.g., links among Web pages

> - Vertices are created one at a time, denoted $1, \ldots, N_v$
> - When node j is created, it makes a single arc to i, $1 \leq i < j$
> - Creation of (j, i) governed by a probabilistic rule:
>   - With probability p, j links to i chosen uniformly at random
>   - With probability $1 - p$, j links to i with probability $\propto d_i^{in}$

- The resulting graph is directed, each vertex has $d_v^{out} = 1$
- Preferential attachment model leads to "rich-gets-richer" dynamics

  $\Rightarrow$ Arcs formed preferentially to (currently) most popular nodes

  $\Rightarrow$ Prob. that i increases its popularity $\propto$ i's current popularity

Preferential attachment yields power laws

Theorem

The preferential attachment model gives rise to a power-law in-degree distribution with exponent $\alpha = 1 + \frac{1}{1-p}$, *i.e.,*

$$P\left(d^{in} = d\right) \propto d^{-\left(1+\frac{1}{1-p}\right)}$$

- Key: "j links to i with probability $\propto d_i^{in}$" equivalent to copying, i.e., "j chooses k uniformly at random, and links to i if $(k, i) \in E$"
- Reflect: Copy other's decision vs. independent decisions in $G_{n,p}$
- As $p \to 0$ $\Rightarrow$ Copying more frequent $\Rightarrow$ Smaller $\alpha \to 2$
  - Intuitive: more likely to see extremely popular pages (heavier tail)

Continuous approximation

- In-degree $d_i^{in}(t)$ of node i at time $t \geq i$ is a RV. Two facts
  F1) Initial condition: $d_i^{in}(i) = 0$ since node i just created at time $t = i$
  F2) Dynamics of $d_i^{in}(t)$: Probability that new node $t + 1 > i$ links to i is

$$\mathrm{P}\left((t+1, i) \in E\right) = p \times \frac{1}{t} + (1-p) \times \frac{d_i^{in}(t)}{t}$$

- Will study a deterministic, continuous approximation to the model
  - Continuous time $t \in [0, N_v]$
  - Continuous degrees $x_i^{in}(t) : [i, N_v] \mapsto \mathbb{R}_+$ are deterministic

- Require in-degrees to satisfy the following growth equation

$$\frac{dx_i^{in}(t)}{dt} = \frac{p}{t} + \frac{(1-p)x_i^{in}(t)}{t}, \quad x_i^{in}(i) = 0$$

Solving the differential equation

- Solve the first-order differential equation for $x_i^{in}(t)$ (let $q = 1 - p$)

$$\frac{dx_i^{in}}{dt} = \frac{p + qx_i^{in}}{t}$$

- Divide both sides by $p + qx_i^{in}(t)$ and integrate over t

$$\int \frac{1}{p + qx_i^{in}} \frac{dx_i^{in}}{dt} dt = \int \frac{1}{t} dt$$

- Solving the integrals, we obtain (c is a constant)

$$\ln\left(p + qx_i^{in}\right) = q \ln\left(t\right) + c$$

Solving the differential equation (cont.)

- Exponentiating and letting $K = e^c$ we find

$$\ln\left(p + qx_i^{in}(t)\right) = q\ln(t) + c \;\Rightarrow\; x_i^{in}(t) = \frac{1}{q}\left(Kt^q - p\right)$$

- To determine the unknown constant K, use the initial condition

$$0 = x_i^{in}(i) = \frac{1}{q}\left(Ki^q - p\right) \;\Rightarrow\; K = \frac{p}{i^q}$$

- Hence, the deterministic approximation of $d_i^{in}(t)$ evolves as

$$x_i^{in}(t) = \frac{1}{q}\left(\frac{p}{i^q} \times t^q - p\right) = \frac{p}{q}\left[\left(\frac{t}{i}\right)^q - 1\right]$$

Obtaining the degree distribution

- Q: At time t, what fraction $\bar{F}(d)$ of all nodes have in-degree $\geq d$?
  Approximation: What fraction of all functions $x_i^{in}(t) \geq d$ by time t?

$$x_i^{in}(t) = \frac{p}{q}\left[\left(\frac{t}{i}\right)^q - 1\right] \geq d$$

- Can be rewritten in terms of i as

$$i \leq t\left[\left(\frac{q}{p}\right)d + 1\right]^{-1/q}$$

- By time t there are exactly t nodes in the graph, so the fraction is

$$\bar{F}(d) = \left[\left(\frac{q}{p}\right)d + 1\right]^{-1/q} = 1 - F(d)$$

Identifying the power law

- The degree distribution is given by the PDF $p(d)$
- Recall that the PDF, CDF and CCDF are related, namely

$$p(x) = \frac{dF(x)}{dx} = -\frac{d\bar{F}(x)}{dx}$$

- Differentiating $\bar{F}(d) = \left[\left(\frac{q}{p}\right) d + 1\right]^{-1/q}$ yields

$$p(d) = \frac{1}{p}\left[\left(\frac{q}{p}\right) d + 1\right]^{-(1+\frac{1}{q})}$$

- Showed $p(d) \propto d^{-(1+1/q)}$, a power law with exponent $\alpha = 1 + \frac{1}{1-p}$
  - $\Rightarrow$ Disclaimer: Relied on heuristic arguments
  - $\Rightarrow$ Rigorous, probabilistic analysis possible

The Barabási-Albert model

- Barabási-Albert (BA) model is for undirected graphs

> - Initial graph $G_{BA}(0)$ of $N_v(0)$ vertices and $N_e(0)$ edges $(t = 0)$
> - For $t = 1, 2, \ldots$ current graph $G_{BA}(t-1)$ grows to $G_{BA}(t)$ by:
>   - Adding a new vertex u of degree $d_u(t) = m \geq 1$
>   - The new edges are incident to m different vertices in $G_{BA}(t-1)$
>   - New vertex u is connected to $v \in \mathcal{V}(t-1)$ w.p.
>
> $$\mathrm{P}\left((u, v) \in \mathcal{E}(t)\right) = \frac{d_v(t-1)}{\sum_{v'} d_{v'}(t-1)}$$

- Vertices connected to u preferentially towards higher degrees

$\Rightarrow$ $G_{BA}(t)$ has $N_v(t) = N_v(0) + t$ and $N_e(t) = N_e(0) + tm$

A. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509-512, 1999

Linearized chord diagram

- BA model ambiguous in how to select m vertices $\propto$ to their degree
  $\Rightarrow$ Joint distribution not specified by marginal on each vertex

- Linearzied chord diagram (LCD) model removes ambiguities

- For $m = 1$, $G_{LCD}(0)$ consists of a vertex with a self-loop
- For $t = 1, 2, \ldots$ graph $G_{LCD}(t-1)$ grows to $G_{LCD}(t)$ by:
  - Adding a new vertex v_t with an edge to $v_s \in \mathcal{V}(t)$
  - Vertex v_s, $1 \leq s \leq t$ is chosen w.p.

$$\mathrm{P}\left(s = j\right) = \begin{cases} \frac{d_{v_j}(t-1)}{2t-1}, & \text{if } 1 \leq j \leq t-1, \\ \frac{1}{2t-1}, & \text{if } j = t \end{cases}$$

- For $m > 1$ simply run the above process m times for each t
  - Collapse all created vertices into a single one, retaining edges

A. Bollobás et al, "The degree sequence of a scale-free random graph process," *Random Struct. and Alg.*, vol. 18, pp. 279-290, 2001

Properties of the LCD model

P1) The LCD model allows for loops and multi-edges, occurring rarely

P2) $G_{LCD}(t)$ has power-law degree distribution with $\alpha = 3$, as $t \to \infty$

P3) The BA model yields connected graphs if $G_{BA}(0)$ connected

$\Rightarrow$ Not true for the LCD model, but $G_{LCD}(t)$ connected w.h.p.

P4) Small-world behavior

$$\text{diam}(G_{LCD}(t)) = \begin{cases} O(\log N_v(t)), & m = 1 \\ O\left(\frac{\log N_v(t)}{\log \log N_v(t)}\right), & m > 1 \end{cases}$$

P5) Unsatisfactory clustering, since small for $m > 1$

$$\mathbb{E}\left[\text{cl}(G_{LCD}(t))\right] \approx \frac{m-1}{8} \frac{(\log N_v(t))^2}{N_v(t)}$$

$\Rightarrow$ Marginally better than $O(N_v^{-1})$ in classical random graphs

Copying models

- Copying is another mechanism of fundamental interest
  Ex: gene duplication to re-use information in organism's evolution
- Different from preferential attachment, but still results in power laws

> - Initialize with a graph $G_C(0)$ $(t = 0)$
> - For $t = 1, 2, \ldots$ current graph $G_C(t-1)$ grows to $G_C(t)$ by:
>   - Adding a new vertex u
>   - Choosing $v \in \mathcal{V}(t-1)$ with uniform probability $\frac{1}{N_v(t-1)}$
>   - Joining vertex u with v's neighbors independently w.p. p

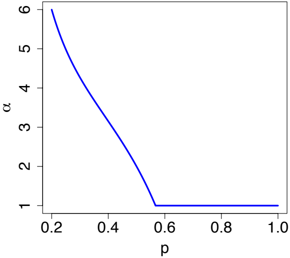
- Case $p = 1$ leads to full duplication of edges from an existing node

F. Chung et al, "Duplication models for biological networks," *Journal of Computational Biology*, vol. 10, pp. 677-687, 2003

Asymptotic degree distribution

- Degree distribution tends to a power law w.h.p. [Chung et al'03]
  - $\Rightarrow$ Exponent α is the plotted solution to the equation

$$p(\alpha - 1) = 1 - p^{\alpha - 1}$$



- Full duplication does not lead to power-law behavior; but does if
  - $\Rightarrow$ Partial duplication performed a fraction $q \in (0, 1)$ of times

Fitting network growth models

- Most common practical usage of network growth models is predictive
  Goal: compare characteristics of G^{obs} and $G(t)$ from the models
- Little attempt to date to fit network growth models to data
  $\Rightarrow$ Expected due to simplicity of such models
  $\Rightarrow$ Still useful to estimate e.g., the duplication probability p
- To fit a model ideally would like to observe a sequence $\{G^{obs}(\tau)\}_{\tau=1}^{t}$
  $\Rightarrow$ Unfortunately, such dynamic network data is still fairly elusive
- **Q:** Can we fit a network growth model to a single snap-shot G^{obs}?
- **A:** Yes, if we leverage the Markovianity of the growth process

Duplication-attachment models

- Similar to all network growth models described so far, suppose:

  As1: A single vertex is added to $G(t-1)$ to create $G(t)$; and
  As2: The manner in which it is added depends only on $G(t-1)$

- In other words, we assume $\{G(t)\}_{t=0}^{\infty}$ is a Markov chain

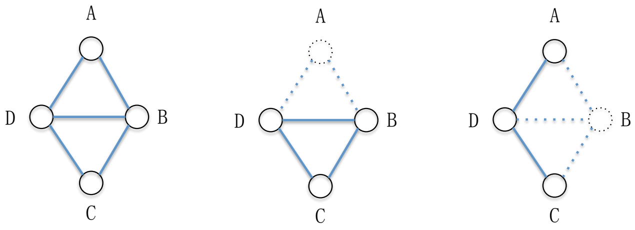
- Let graph $\delta(G(t), v)$ be obtained by deleting v and its edges from $G(t)$

- **Def:** vertex v is removable if $G(t)$ can be obtained from $\delta(G(t), v)$ via copying. If $G(t)$ has no removable vertices, we call it irreducible

- The class of duplication-attachment (DA) models satisfies:
  (i) The initial graph $G(0)$ is irreducible; and
  (ii) $\mathrm{P}_{\theta}(G(t) \,|\, G(t-1)) > 0 \Leftrightarrow G(t)$ obtained by copying a vertex in $G(t-1)$

C. Wiuf et al, "A likelihood approach to analysis of network data," *PNAS*, vol. 103, pp. 7566-7570, 2006

Example: reducible graph



$G(t)$ $\delta(G(t), v_A)$ $\delta(G(t), v_B)$

- Vertex v_A is removable (likewise v_c by symmetry)
  - $\Rightarrow$ Obtain $G(t)$ from $\delta(G(t, v_a))$ by copying v_c
- This implies that $G(t)$ is reducible
  - $\Rightarrow$ Notice though that v_B or v_D are not removable

MLE for DA model parameters

- Suppose that $G^{obs} = G(t)$ represents the observed network graph
- The likelihood for the parameter θ is recursively given by

$$\mathcal{L}(\theta; G(t)) = \frac{1}{t} \sum_{v \in \mathcal{R}_{G(t)}} \mathrm{P}_\theta \left(G(t) \,\middle|\, \delta(G(t), v) \right) \mathcal{L}(\theta; \delta(G(t), v))$$

$\Rightarrow$ $\mathcal{R}_{G(t)}$ is the set of all removable nodes in $G(t)$

- The MLE for θ is thus defined as

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta; G(t))$$

$\Rightarrow$ Computing $\mathcal{L}(\theta; G(t))$ non-trivial, even for modest-size graphs

- Monte Carlo methods to approximate $\mathcal{L}(\theta; G(t))$ [Wiuf et al'06]

$\Rightarrow$ Open issues: vector $\boldsymbol{\theta}$, other growth models, scalability

Roadmap

Random graph models

Small-world models

Network-growth models

Exponential random graph models

Latent network models

Random dot product graphs

Statistical network graph models

- Good statistical network graph models should be [Robbins-Morris'07]:
  - ⇒ Estimable from and reasonably representative of the data
  - ⇒ Theoretically plausible about the underlying network effects
  - ⇒ Discriminative among competing effects to best explain the data
- Network-based versions of canonical statistical models
  - ⇒ Regression models - Exponential random graph models (ERGMs)
  - ⇒ Latent variable models - Latent network models
  - ⇒ Mixture models - Stochastic block models
- Focus here on ERGMs, also known as p^* models

G. Robbins et al., "An introduction to exponential random graph (p^*) models for social networks," *Social Networks*, vol. 29, pp. 173-191, 2007

Exponential family

- **Def:** discrete random vector $\mathbf{Z} \in \mathcal{Z}$ belongs to an exponential family if

$$\mathrm{P}_\theta(\mathbf{Z} = \mathbf{z}) = \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{z}) - \psi(\theta)\right\}$$

  - $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of parameters and $\mathbf{g} : \mathcal{Z} \mapsto \mathbb{R}^p$ is a function
  - $\psi(\theta)$ is a normalization term, ensuring $\sum_{\mathbf{z} \in \mathcal{Z}} \mathrm{P}_\theta(\mathbf{z}) = 1$
  - Ex: Bernoulli, binomial, Poisson, geometric distributions

- For continuous exponential families, the pdf has an analogous form

  Ex: Gaussian, Pareto, chi-square distributions

- Exponential families share useful algebraic and geometric properties

  $\Rightarrow$ Mathematically convenient for inference and simulation

Exponential random graph model

- Let $G(\mathcal{V}, \mathcal{E})$ be a random undirected graph, with $A_{ij} := \mathbb{I}\{(i,j) \in \mathcal{E}\}$
  - Matrix $\mathbf{A} = [A_{ij}]$ is the random adjacency matrix, $\mathbf{a}$ a realization
- An ERGM specifies in exponential family form the distribution of $\mathbf{A}$, i.e.,

$$\mathrm{P}_\theta(\mathbf{A} = \mathbf{a}) = \left(\frac{1}{\kappa(\boldsymbol{\theta})}\right) \exp\left\{\sum_H \theta_H g_H(\mathbf{a})\right\}, \quad \text{where}$$

(i) each H is a configuration, meaning a set of possible edges in G;
(ii) $g_H(\mathbf{a})$ is the network statistic corresponding to configuration H

$$g_H(\mathbf{a}) = \prod_{A_{ij} \in H} A_{ij} = \mathbb{I}\{H \text{ occurs in } \mathbf{a}\}$$

(iii) $\theta_H \neq 0$ only if all edges in H are conditionally dependent; and
(iv) $\kappa(\boldsymbol{\theta})$ is a normalization constant ensuring $\sum_{\mathbf{a}} \mathrm{P}_\theta(\mathbf{a}) = 1$

Discussion

- Graph order N_v is fixed and given, only edges are random

  $\Rightarrow$ Assumed unweighted, undirected edges. Extensions possible

- ERGMs describe random graphs 'built-on' localized patterns
  - These configurations are the structural characteristics of interest
  - Ex: Are there reciprocity effects? Add mutual arcs as configurations
  - Ex: Are there transitivity effects? Consider triangles

- (In)dependence is conditional on all other variables (edges) in G

  $\Rightarrow$ Control configurations relevant (i.e., $\theta_H \neq 0$) to the model

- Well-specified dependence assumptions imply particular model classes

A general framework for model construction

- In positing an ERGM for a network, one implicitly follows five steps

  $\Rightarrow$ Explicit choices connecting hypothesized theory to data analysis

  Step 1: Each edge (relational tie) is regarded as a random variable

  Step 2: A dependence hypothesis is proposed

  Step 3: Dependence hypothesis implies a particular form to the model

  Step 4: Simplification of parameters through e.g., homogeneity

  Step 5: Estimate and interpret model parameters

Example: Bernoulli random graphs

- Assume edges present independently of all other edges (e.g., in $G_{n,p}$)
  $\Rightarrow$ Simplest possible (and unrealistic) dependence assumption
- For each (i, j), we assume $_{ij}$ independent of A_{uv}, for all $(u, v) \neq (i, j)$
  $\Rightarrow \theta_H = 0$ for all H involving two or more edges
- Edge configurations i.e., $g_H(\mathbf{a}) = A_{ij}$ relevant, and the ERGM becomes

$$\mathrm{P}_\theta(\mathbf{A} = \mathbf{a}) = \left(\frac{1}{\kappa(\boldsymbol{\theta})}\right) \exp\left\{\sum_{i,j} \theta_{ij} A_{ij}\right\}$$

- Specifies that edge (i, j) present independently, with probability

$$p_{ij} = \frac{\exp(\theta_{ij})}{1 + \exp(\theta_{ij})}$$

Constraints on parameters: homogeneity

- Too many parameters makes estimation infeasible from single **a**
  - $\Rightarrow$ Under independence have N_v^2 parameters $\{\theta_{ij}\}$. Reduction?
- Homogeneity across all G, i.e., $\theta_{ij} = \theta$ for all (i, j) yields

$$\mathrm{P}_\theta(\mathbf{A} = \mathbf{a}) = \left(\frac{1}{\kappa(\boldsymbol{\theta})}\right) \exp\left\{\theta L(\mathbf{a})\right\}$$

  - Relevant statistic is the number of edges observed $L(\mathbf{a}) = \sum_{i,j} A_{ij}$
  - ERGM identical to $G_{n,p}$, where $p = \dfrac{\exp\theta}{1 + \exp\theta}$

Ex: suppose we know a priori that vertices fall in two sets

- Can impose homogeneity on edges within and between sets, i.e.,

$$\mathrm{P}_\theta(\mathbf{A} = \mathbf{a}) = \left(\frac{1}{\kappa(\boldsymbol{\theta})}\right) \exp\left\{\theta_1 L_1(\mathbf{a}) + \theta_{12} L_{12}(\mathbf{a}) + \theta_2 L_2(\mathbf{a})\right\}$$

Example: Markov random graphs

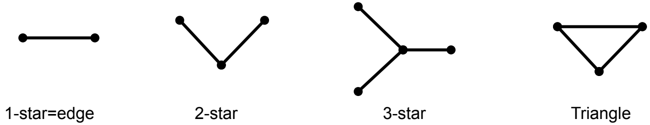
- Markov dependence notion for network graphs [Frank-Strauss'86]
  - Assumes two ties are dependent if they share a common node
  - Edge status A_{ij} dependent on any other edge involving i or j

Theorem

Under homogeneity, G is a Markov random graph if and only if

$$P_\theta(\mathbf{A} = \mathbf{a}) = \left(\frac{1}{\kappa(\boldsymbol{\theta})}\right) \exp\left\{\sum_{k=1}^{N_v-1} \theta_k S_k(\mathbf{a}) + \theta_\tau T(\mathbf{a})\right\}, \quad \textit{where}$$

$S_k(\mathbf{a})$ is the number of k-stars, and $T(\mathbf{a})$ the number of triangles



Alternative statistics

- Including many higher-order terms challenges estimation
  - $\Rightarrow$ High-order star effects often omitted, e.g., $\theta_k = 0$, $k \geq 4$
  - $\Rightarrow$ But these models tend to fit real data poorly. Dilemma?
- Idea: Impose parametric form $\theta_k \propto (-1)^k \lambda^{2-k}$ [Snijders et al'06]
- Combine $S_k(\mathbf{a})$, $k \geq 2$ into a single alternating k-star statistic, i.e.,

$$\text{AKS}_\lambda(\mathbf{a}) = \sum_{k=2}^{N_v-1} (-1)^k \frac{S_k(\mathbf{a})}{\lambda^{k-2}}, \quad \lambda > 1$$

- Can show $\text{AKS}_\lambda(\mathbf{a}) \propto$ the geometrically-weighted degree count

$$\text{GWD}_\gamma(\mathbf{a}) = \sum_{d=0}^{N_v-1} e^{-\gamma d} N_d(\mathbf{a}), \quad \gamma > 0$$

  - $\Rightarrow$ $N_d(\mathbf{a})$ is the number of vertices with degree d

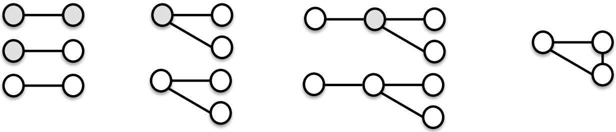
Incorporating vertex attributes

- Straightforward to incorporate vertex attributes to ERGMs
  Ex: gender, seniority in organization, protein function
- Consider a realization $\mathbf{x}$ of a random vector $\mathbf{X} \in \mathbb{R}^{N_v}$ defined on $\mathcal{V}$
- Specify an exponential family form for the conditional distribution

$$\mathrm{P}_\theta(\mathbf{A} = \mathbf{a} \,\big|\, \mathbf{X} = \mathbf{x})$$

$\Rightarrow$ Will include additional statistics $g(\cdot)$ of $\mathbf{a}$ and $\mathbf{x}$

- Ex: configurations for Markov, binary vertex attributes



Estimating ERGM parameters

- MLE for the parameter vector $\boldsymbol{\theta}$ in an ERGM is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a}) - \psi(\boldsymbol{\theta}) \right\}, \quad \text{where } \psi(\boldsymbol{\theta}) := \log \kappa(\boldsymbol{\theta})$$

- Optimality condition yields

$$\mathbf{g}(\mathbf{a}) = \nabla \psi(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

- Using also that $\mathbb{E}_\theta[\mathbf{g}(\mathbf{A})] = \nabla\psi(\boldsymbol{\theta})$, the MLE solves

$$\mathbb{E}_{\hat{\theta}}[\mathbf{g}(\mathbf{A})] = \mathbf{g}(\mathbf{a})$$

- Unfortunately $\psi(\boldsymbol{\theta})$ cannot be computed except for small graphs

  $\Rightarrow$ Involves a summation over $2^{\binom{N_v}{2}}$ values of $\mathbf{a}$ for each $\boldsymbol{\theta}$

  $\Rightarrow$ Numerical methods needed to obtain approximate values of $\hat{\boldsymbol{\theta}}$

Proof of $\mathbb{E}\left[g(\mathbf{A})\right] = \nabla\psi(\theta)$

- The pmf of $\mathbf{A}$ is $\mathrm{P}_\theta(\mathbf{A} = \mathbf{a}) = \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a}) - \psi(\theta)\right\}$, hence

$$\begin{aligned}\mathbb{E}_\theta[g(\mathbf{A})] &= \sum_{\mathbf{a}} g(\mathbf{a})\mathrm{P}_\theta(\mathbf{A} = \mathbf{a}) \\ &= {\color{red}\sum_{\mathbf{a}} g(\mathbf{a}) \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a}) - \psi(\theta)\right\}}\end{aligned}$$

- Recall $\psi(\boldsymbol{\theta}) = \log \sum_{\mathbf{a}} \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a})\right\}$ and use the chain rule

$$\begin{aligned}\nabla\psi(\boldsymbol{\theta}) &= \frac{\sum_{\mathbf{a}} g(\mathbf{a}) \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a})\right\}}{\sum_{\mathbf{a}} \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a})\right\}} = \frac{\sum_{\mathbf{a}} g(\mathbf{a}) \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a})\right\}}{\exp \psi(\theta)} \\ &= {\color{blue}\sum_{\mathbf{a}} g(\mathbf{a}) \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a}) - \psi(\theta)\right\}}\end{aligned}$$

- The red and blue sums are identical $\Rightarrow \mathbb{E}_\theta[g(\mathbf{A})] = \nabla\psi(\boldsymbol{\theta})$ follows

Markov chain Monte Carlo MLE

- Idea: for fixed $\boldsymbol{\theta}_0$, maximize instead the log-likelihood ratio

$$r(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0) = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{g}(\mathbf{a}) - [\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)]$$

- Key identity: will show that

$$\exp\{\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)\} = \mathbb{E}_{\theta_\mathbf{0}}\left[\exp\left\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{g}(\mathbf{A})\right\}\right]$$

- Markov chain Monte Carlo MLE algorithm to search over $\boldsymbol{\theta}$

  Step 1: draw samples $\mathbf{A}_1, \ldots, \mathbf{A}_n$ from the ERGM under $\boldsymbol{\theta}_0$

  Step 2: approximate the above $\mathbb{E}_{\theta_\mathbf{0}}[\cdot]$ via sample averaging

  Step 3: the logarithm of the result approximates $\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)$

  Step 4: evaluate an $\approx$ log-likelihood ratio $r(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$

- For large n, the maximum value found approximates the MLE $\hat{\boldsymbol{\theta}}$

Derivation of key identity

- Recall $\exp \psi(\boldsymbol{\theta}) = \sum_{\mathbf{a}} \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a})\right\}$ to write

$$\exp\left\{\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)\right\} = \frac{\sum_{\mathbf{a}} \exp\left\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{a})\right\}}{\exp \psi(\boldsymbol{\theta}_0)}$$

- Multiplying and dividing by $\exp\left\{\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{a})\right\} > 0$ yields

$$\begin{aligned}
\exp\left\{\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)\right\} &= \sum_{\mathbf{a}} \exp\left\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{g}(\mathbf{a})\right\} \times \frac{\exp\left\{\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{a})\right\}}{\exp \psi(\boldsymbol{\theta}_0)} \\
&= \sum_{\mathbf{a}} \exp\left\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{g}(\mathbf{a})\right\} \mathrm{P}_{\theta_0}(\mathbf{A} = \mathbf{a}) \\
&= \mathbb{E}_{\theta_0}\left[\exp\left\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{g}(\mathbf{A})\right\}\right]
\end{aligned}$$

- Used $\exp\left\{\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{a}) - \psi(\boldsymbol{\theta}_0)\right\}$ is the exponential family pmf $\mathrm{P}_{\theta_0}(\mathbf{A} = \mathbf{a})$

Model goodness-of-fit

- Best fit chosen from a given class of models ...
  may not be a good fit to the data if model class not rich enough

- Assessing goodness-of-fit for ERGMs

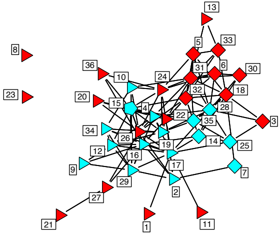
  Step 1: simulate numerous random graphs from the fitted model
  Step 2: compare high-level characteristics with those of G^{obs}
  Ex: distributions of degree, centrality, diameter

- If significant differences found in G^{obs}, conclude
  - $\Rightarrow$ Systematic gap between specified model class and data
  - $\Rightarrow$ Lack of goodness-of-fit

- **Take home:** model specification for ERGMs highly nontrivial
  - $\Rightarrow$ Goodness-of-fit diagnostics can play key facilitating role

Example: Lawyer collaboration network

- Network G^{obs} of working relationships among lawyers [Lazega'01]
  - Nodes are $N_v = 36$ partners, edges indicate partners worked together



- Data includes various node-level attributes:
  - Seniority (node labels indicate rank ordering)
  - Office location (triangle, square or pentagon)
  - Type of practice, i.e., litigation (red) and corporate (cyan)
  - Gender (three partners are female labeled 27, 29 and 34)

- Goal: study cooperation among social actors in an organization

Modeling lawyer collaborations

- Assess network effects $S_1(\mathbf{a}) = N_e$ and alternating k-triangles statistic

$$\text{AKT}_\lambda(\mathbf{a}) = 3T_1(\mathbf{a}) + \sum_{k=2}^{N_v-2} (-1)^{k+1} \frac{T_k(\mathbf{a})}{\lambda^{k-1}}$$

  $\Rightarrow$ $T_k(\mathbf{a})$ counts sets of k individual triangles sharing a common base

- Test the following set of exogenous effects:

$$h^{(1)}(\mathbf{x}_i, \mathbf{x}_j) = \text{seniority}_i + \text{seniority}_j, \quad h^{(2)}(\mathbf{x}_i, \mathbf{x}_j) = \text{practice}_i + \text{practice}_j$$
$$h^{(3)}(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{I}\left\{\text{practice}_i = \text{practice}_j\right\}, \quad h^{(4)}(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{I}\left\{\text{gender}_i = \text{gender}_j\right\}$$
$$h^{(5)}(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{I}\left\{\text{office}_i = \text{office}_j\right\}, \quad \mathbf{h}(\mathbf{x}_i, \mathbf{x}_j) := [h^{(1)}(\mathbf{x}_i, \mathbf{x}_j), \ldots, h^{(5)}(\mathbf{x}_i, \mathbf{x}_j)]^T$$

- Resulting ERGM

$$\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\beta}}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{1}{\kappa(\boldsymbol{\theta}, \boldsymbol{\beta})} \exp\left\{\theta_1 S_1(\mathbf{a}) + \theta_2 \text{AKT}_\lambda(\mathbf{a}) + \boldsymbol{\beta}^T \mathbf{g}(\mathbf{a}, \mathbf{x})\right\}$$
$$\mathbf{g}(\mathbf{a}, \mathbf{x}) = \sum_{i,j} A_{ij} \mathbf{h}(\mathbf{x}_i, \mathbf{x}_j)$$

Model fitting result

- Fitting results using the MCMC MLE approach

Parameter	Estimate	'Standard Error'
Density (θ_1)	-6.2073	0.5697
Alternating k-triangles (θ_2)	0.5909	0.0882
Seniority Main Effect (β_1)	0.0245	0.0064
Practice Main Effect (β_2)	0.3945	0.1103
Same Practice (β_3)	0.7721	0.1973
Same Gender (β_4)	0.7302	0.2495
Same Office (β_5)	1.1614	0.1952

$\Rightarrow$ Standard errors heuristically obtained via asymptotic theory

- Identified factors that may increase odds of cooperation

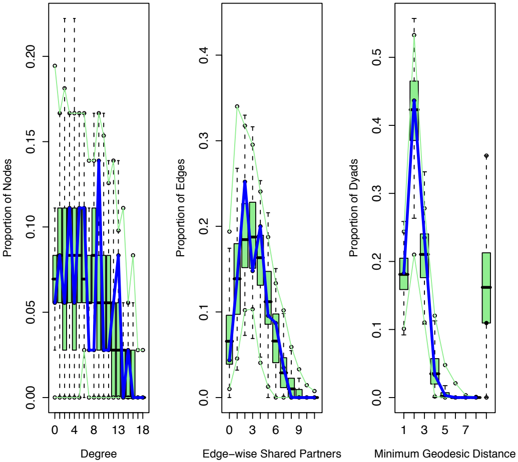
  Ex: same practice, gender and office location double odds

- Strong evidence for transitivity effects since $\hat{\theta}_2 \gg \mathrm{se}(\hat{\theta}_2)$

  $\Rightarrow$ Something beyond basic homophily explaining such effects

Assessing goodness-of-fit

- Assess goodness-of-fit to G^{obs}
  - Sample from fitted ERGM
- Compared distributions of
  - Degree
  - Edge-wise shared partners
  - Geodesic distance
- Plots show good fit overall



Roadmap

Latent network models

- Latent variables widely used to model observed data

  Ex: Hidden Markov models, factor analysis

- Basic idea permeated to statistical network analysis. Two types:

- Latent class models: unobserved class membership drives propensity towards establishing relational ties

- Latent feature models: relational ties more likely to form among vertices that are 'closer' in some latent space

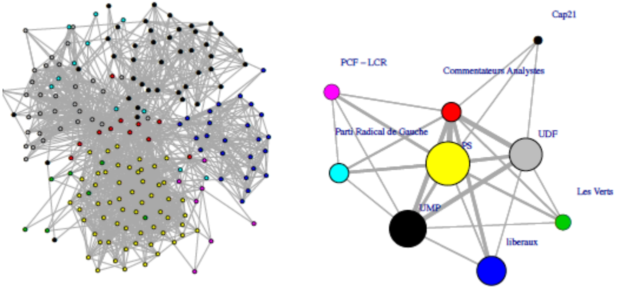
- As of now latent network models come in many flavors. Focus here:

  ⇒ Stochastic block models (SBMs)

  ⇒ More general non-parametric analog based on graphons

Example: French political blogs

- French political blog network from October 2006 [Kolaczyk'17]
  - ⇒ Consists of $N_v = 192$ blogs linked by $N_e = 1431$ edges
  - ⇒ Colors indicate blog affiliation to a French political party



- Visual evidence of mixing of smaller subgraphs
  - ⇒ Different rates of connections among blogs (driven by party)
  - ⇒ Erdös-Renyi with fixed p cannot capture this structure

Stochastic block models

- Stochastic block models explicitly parameterize the notion of
  - $\Rightarrow$ Groups, modules or communities $\mathcal{C}_1, \ldots, \mathcal{C}_Q$
  - $\Rightarrow$ Connection rates π_{qr} of vertices between/within groups

A generative model for an undirected random graph $G(\mathcal{V}, \mathcal{E})$

- Fix Q. Each vertex $i \in \mathcal{V}$ independently belongs to $\mathcal{C}_q$ w.p. α_q

$$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_Q]^\top, \quad \mathbf{1}^\top \boldsymbol{\alpha} = 1$$

- For vertices $i, j \in \mathcal{V}$, with $i \in \mathcal{C}_q$ and $j \in \mathcal{C}_r$ $\Rightarrow (i, j) \in \mathcal{E}$ w.p. π_{qr}

P. W. Holland et al., "Stochastic block-models: First steps," *Social Networks*, vol. 5, pp. 109-137, 1983

Model specification and flexibility

- In other words, with $Z_{iq} = \mathbb{I}\{i \in \mathcal{C}_q\}$ and $\mathbf{Z}_i = [Z_{i1}, \ldots, Z_{iQ}]^\top$

$$\mathbf{Z}_i \overset{\text{i.i.d.}}{\sim} \text{Multinomial}(1, \boldsymbol{\alpha}),$$
$$A_{ij} \,\big|\, \mathbf{Z}_i = \mathbf{z}_i, \mathbf{Z}_j = \mathbf{z}_j \sim \text{Bernoulli}(\pi_{\mathbf{z}_i,\mathbf{z}_j})$$

  for $1 \leq i, j \leq N_v$, where $A_{ij} = A_{ji}$ and $A_{ii} \equiv 0$

- Parameters: Q mixing weights α_q and Q^2 connection probs. π_{qr}

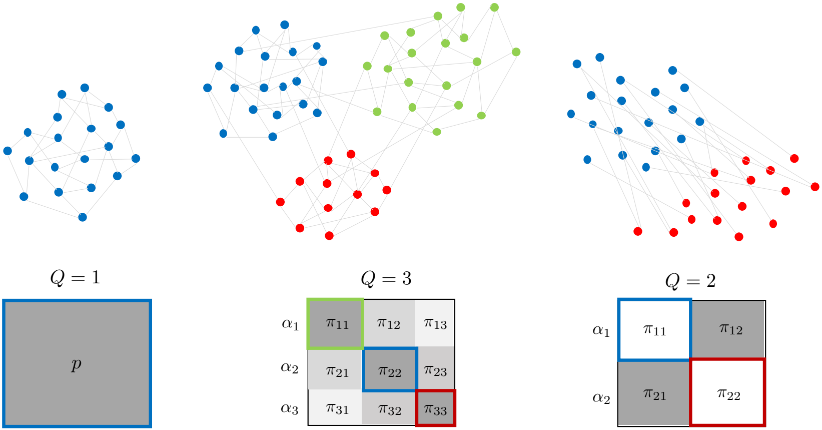
- Mixture of classical random graph models

$$\mathrm{P}\left(A_{ij} = 1\right) = \sum_{1 \leq q,r \leq Q} \alpha_q \alpha_r \pi_{qr}$$

  $\Rightarrow$ More flexible to capture the structure of observed networks

  $\Rightarrow$ May face issues of identifiability [Allman et al'11]

- Emergence of giant component, size distribution of groups [Söderberg'03]



- Mixtures of Erdös-Renyi models can be surprisingly flexible

Graphons and f−random graphs

- A non-parametric variant of the SBM can be defined as follows:

$$U_1, \ldots, U_{N_v} \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,1],$$
$$A_{ij} \,\Big|\, U_i = u_i, U_j = u_j \sim \text{Bernoulli}(f(u_i, u_j))$$

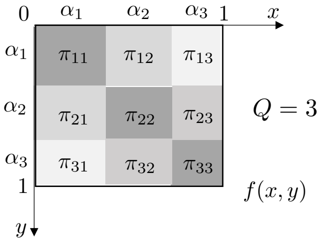
  for $1 \leq i, j \leq N_v$, where $A_{ij} = A_{ji}$ and $A_{ii} \equiv 0$

- Graphon: symmetric, measurable function $f : [0,1]^2 \mapsto [0,1]$
  - $\Rightarrow$ Resulting graph G known as f−random graph
- Latent U_i randomly assigns vertex positions uniformly in $[0,1]$
  - $\Rightarrow$ Graphon $f(u_i, u_j)$ specifies connection rate between i, j
- SBM: Latent $\mathbf{Z}_i$ assigns memberships of vertices to one of Q groups
  - $\Rightarrow$ Probability π_{qr} defines linking rate between $i \in \mathcal{C}_q, j \in \mathcal{C}_r$

L. Lovász, "Large Networks and Graph Limits," *AMS Colloquium Publications*, vol. 60, 2012

Example: SBM graphon

- The $f-$random graph model subsumes the parametric SBM. Recipe
  (i) Partition $[0,1]$ into Q subintervals of lengths $\alpha_1, \ldots, \alpha_Q$
  (ii) Take the Cartesian product to partition $[0,1]^2$ into Q^2 blocks
  (iii) Define f to be piecewise constant on blocks, qrth block has height π_{qr}



- Can approximate measurable functions by piecewise-constant functions

  $\Rightarrow$ Approximate $f-$random graphs (in distribution) with an SBM

  $\Rightarrow$ Number of blocks Q required may be huge!

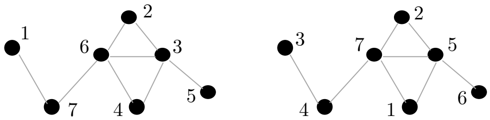
Example: Network generation

- Consider an $f-$random graph with $f(x, y) = \min(x, y)$ [Lovász'12]
  - $\Rightarrow$ Left plot shows a gray-scale rendering of graphon f
- **Q:** How do generated graphs with $N_v = 40$ look like?

- Center plot depicts a realization of the adjacency matrix $\mathbf{A}$
  - $\Rightarrow$ Given $\mathbf{A}$ only, impossible to decipher how the graph was generated
- Sort vertices according to order statistics $U_{(1)}, \ldots, U_{(40)}$ (right plot)
  - $\Rightarrow$ Amenable to graphon estimation via non-parametric regression

Vertex exchangeability

- **Def:** a random array $\mathbf{A} = [A_{ij}]_{i,j \in \mathcal{V}}$ is vertex exchangeable if $\mathbf{A}_\sigma := [A_{\sigma(i)\sigma(j)}]_{i,j \in \mathcal{V}} \overset{D}{=} \mathbf{A}$ for every permutation $\sigma : \mathcal{V} \mapsto \mathcal{V}$



- Exchangeable models assign equal probability to isomorphic graphs
  - $\Rightarrow$ Means said models are most natural for unlabeled graphs
- Like SBMs, one can easily show f-random graphs are exchangeable
- Remarkably, every exchangeable model is a mixture of f-random graphs
  - $\Rightarrow$ Aldous-Hoover theorem extends de Finetti's result for sequences

D. J. Aldous, "Representations for partially exchangeable arrays of random variables," *Journal of Mulivariate Analysis*, vol. 11, 1981

Every f-random graph is exchangeable

- The distribution of an $f-$random graph with N_v vertices is

$$\mathrm{P}(\mathbf{A}=\mathbf{a}) = \int_{[0,1]^{N_v}} \prod_{1\leq i\neq j\leq N_v} f(u_i,u_j)^{A_{ij}}(1-f(u_i,u_j))^{1-A_{ij}} du_1 \ldots du_{N_v}$$

- For arbitrary permutation $\sigma : \mathcal{V} \mapsto \mathcal{V}$ and since the U_i are i.i.d. we have

$$\begin{aligned}
\mathrm{P}(\mathbf{A}_\sigma=\mathbf{a}_\sigma) &= \int_{[0,1]^{N_v}} \prod_{1\leq i\neq j\leq N_v} f(u_i,u_j)^{A_{\sigma(i)\sigma(j)}}(1-f(u_i,u_j))^{1-A_{\sigma(i)\sigma(j)}} du_1 \ldots du_{N_v} \\
&= \int_{[0,1]^{N_v}} \prod_{1\leq i\neq j\leq N_v} f(u_{\sigma^{-1}(i)},u_{\sigma^{-1}(j)})^{A_{ij}}(1-f(u_{\sigma^{-1}(i)},u_{\sigma^{-1}(j)}))^{1-A_{ij}} \\
&\quad \times du_{\sigma^{-1}(1)} \ldots du_{\sigma^{-1}(N_v)} \\
&= \int_{[0,1]^{N_v}} \prod_{1\leq i\neq j\leq N_v} f(u_i,u_j)^{A_{ij}}(1-f(u_i,u_j))^{1-A_{ij}} du_1 \ldots du_{N_v} \\
&= \mathrm{P}(\mathbf{A}=\mathbf{a})
\end{aligned}$$

Identifiability issues

- Parametrization of $f-$random graphs is not unique $\Rightarrow$ Non-identifiable

  Ex: graphons $f(x, y)$ and $f(1-x, 1-y)$ yield the same model since

  $$U \stackrel{D}{=} 1 - U \text{ for } U \sim \text{Uniform}[0,1]$$

  Ex: graphons $f(x, y)$ and $f(\phi(x), \phi(y))$ for measure-preserving ϕ, i.e.,

  $$\phi : [0,1] \mapsto [0,1] \text{ for which } \phi(U) \sim \text{Uniform}[0,1]$$

- Restrictions on the graphon f are often needed for statistical modeling
- **Def:** f is strictly monotone if $\exists\, \phi$ such that $\tilde{f}(x, y) = f(\phi(x), \phi(y))$ has a strictly increasing degree function $\tilde{g}(x) = \int_{[0,1]} \tilde{f}(x, y) dy$

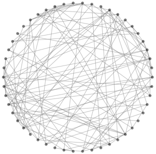
  $\Rightarrow$ Restriction to $\tilde{f}$ yields model identifiability [Bickel-Chen'09]

Graph limits

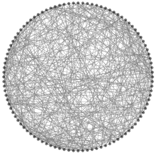
- Graph sequence $G_n(\mathcal{V}_n, \mathcal{E}_n)$ with growing number of nodes $N_v = n$
  - **Q:** When can we say $\{G_n\}_{n=1}^{\infty}$ converges to a limit?
  - **Q:** In what sense is convergence meaningful?
  - **Q:** What kind of object is this limit?

- Spoiler: If the sequence $\{G_n\}_{n=1}^{\infty}$ converges, the limit is a graphon f

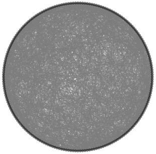
  Ex: sequence of $G_{n,p}$ graphs as $n \to \infty$



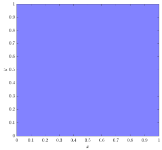
$n = 50$ nodes $\rightarrow$



$n = 100$ nodes $\rightarrow$



$n = 200$ nodes $\rightarrow$



Graphon $f(x, y) \equiv p$

Homomorphism density

- **Def:** Homomorphisms h are adjacency preserving maps from motif $F(\mathcal{V}', \mathcal{E}')$ into graph $G(\mathcal{V}, \mathcal{E})$

$$h : \mathcal{V}' \mapsto \mathcal{V} \text{ such that } (i, j) \in \mathcal{E}' \text{ implies } (h(i), h(j)) \in \mathcal{E}$$

- **Def:** Homomorphism density of motif F in graph G is

$$t(F, G) = \frac{\hom(F, G)}{|\mathcal{V}|^{|\mathcal{V}'|}}$$

  - $\hom(F, G)$: number of homomorphisms between F and G
  - $|\mathcal{V}|^{|\mathcal{V}'|}$: number of maps between vertices in F and G

- Relative measure of the number of ways in which F can be mapped to G

Convergence of graph sequences

- **Def:** A sequence $\{G_n\}_{n=1}^{\infty}$ of graphs converges when for every motif F, the homomorphism density sequence $\{t(F, G_n)\}_{n=1}^{\infty}$ converges

- Noteworthy points about the definition
  - If the sequence becomes constant, then it converges
  - Sequence of isomorphic graphs trivially converges
  - Normalized densities converge, not number edges, triangles, ...
  - Result is for sequence of dense graphs, i.e., $|\mathcal{E}_n| = \Omega(n^2)$

- Answered the first two questions. Need to address the third

  $\Rightarrow$ The limit of a sequence of graphs is not necessarily a graph

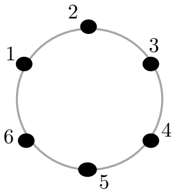
  $\Rightarrow$ **Q:** What kind of object is this limit?

L. Lovász and B. Szegedy, "Limits of dense graph sequences," *Journal of Combinatorial Theory, Series B*, vol. 96, 2006

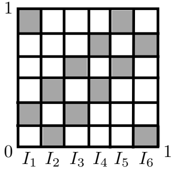
Induced graphon

- Every graph admits a graphon representation termed induced graphon
- Consider a graph $G(\mathcal{V},\mathcal{E})$ with adjacency matrix $\mathbf{A}$
- Uniform partition of $[0,1]$ in N_v subintervals $\Rightarrow I_i = [\frac{i-1}{N_v}, \frac{i}{N_v})$
- **Def:** The induced graphon f_G of G is

$$f_G(x,y) = \sum_{1 \le i,j \le N_v} A_{ij}\mathbb{I}\{x \in I_i\}\mathbb{I}\{y \in I_j\}$$



Cycle graph G with $N_v = 6$ nodes



Graphon f_G induced by the graph G

Limit object is a graphon

- **Claim:** Homomorphism density of motif F in graph G given by

$$t(F, G) = \int_{[0,1]^{|\mathcal{V}'|}} \prod_{(i,j)\in\mathcal{E}'} f_G(u_i, u_j) du_1 \ldots du_{|\mathcal{V}'|}$$

$\Rightarrow$ Probability of F being mapped to an f_G-random graph

- This carries over to the limit. If the sequence $\{G_n\}_{n=1}^{\infty}$ converges, then

$$\lim_{n\to\infty} t(F, G_n) = \int_{[0,1]^{|\mathcal{V}'|}} \prod_{(i,j)\in\mathcal{E}'} f(u_i, u_j) du_1 \ldots du_{|\mathcal{V}'|}$$

for some symmetric, measurable function $f : [0,1]^2 \mapsto [0,1]$

- We identify the limiting object – termed graphon – with f

Why is this useful at all?

Mathematical impact

- Bring to bear analysis tools in an otherwise purely combinatorial context

Statistical inference impact

- Large realizations become representative of the generative process
  - $\Rightarrow$ Infer the data-generation mechanism by examining the realization

Machine learning impact

- Study graph filters and GNNs in the limit of large number of nodes
  - $\Rightarrow$ Transferability e.g., using a trained model on a larger graph

L. Ruiz et al, "Graphon neural networks and the transferability of graph neural networks," *NeurIPS*, 2020

Plausibility

- Good statistical network graph models should be [Robbins-Morris'07]:
  - ⇒ Estimable from and reasonably representative of the data
  - ⇒ Theoretically plausible about the underlying network effects
- **Q:** How appropriate are latent network models? Are they plausible?
- **Q:** Can we approximate well an observed graph G^{obs} with an SBM?
  - ⇒ A variant of the Szemerédi regularity lemma useful here

C. Borgs et al, "Graph limits and parameter testing," *Symposium on Theory of Computing*, 2006

Cut distance

- Discussing approximation notions requires a distance between graphs
- **Def:** For graphs $G(\mathcal{V}, \mathcal{E})$ and $G'(\mathcal{V}', \mathcal{E}')$ with $|\mathcal{V}| = |\mathcal{V}'| = N_v$, the cut distance is given by

$$d_{\square}(G, G') = \frac{1}{N_v^2} \max_{\mathcal{S}, \mathcal{T} \in \{1, \ldots, N_v\}} \Big| \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{T}} (A_{ij} - A'_{ij}) \Big|$$

$\Rightarrow$ One can show the quantity $d_{\square}(\cdot, \cdot)$ is a formal metric

- Defining and studying properties of graph distances is a timely topic

B. Bollobás and O. Riordan, "Sparse graphs: Metrics and random models," *Random Structures & Algorithms*, vol. 39, 2011

An approximation result

- Let $\mathcal{P} = \{\mathcal{V}_1, \ldots, \mathcal{V}_Q\}$ partition the vertices $\mathcal{V}$ of G into Q classes
- Define the complete graph G_P with vertex set $\mathcal{V}$ and edge weights

$$w_{ij}(G_P) = \frac{1}{|\mathcal{V}_q||\mathcal{V}_r|} \sum_{u \in \mathcal{V}_q} \sum_{v \in \mathcal{V}_r} A_{uv}, \quad i \in \mathcal{V}_q, j \in \mathcal{V}_r$$

$\Rightarrow$ Expectation of a $Q-$class block model approximation to G

$\Rightarrow$ Probability an edge joins i, j is just $w_{ij}(G_P)$

Theorem: For every $\epsilon > 0$ and every graph $G(\mathcal{V}, \mathcal{E})$, there exists a partition $\mathcal{P}$ of $\mathcal{V}$ into $Q \leq 2^{\frac{2}{\epsilon^2}}$ classes such that $d_{\square}(G, G_P) \leq \epsilon$.

- Justifies the claim that an SBM can approximate well an arbitrary graph

$\Rightarrow$ The upper bound on Q can be prohibitively large

What about f-random graphs?

- The $f-$random graph model is only appropriate for dense networks

Theorem: If a graph G is the restriction to vertices $\{1, \ldots, N_v\}$ of an infinite exchangeable random graph, then it is either dense or empty.

Proof sketch: The expected proportion of edges in $G(\mathcal{V}, \mathcal{E})$ is

$$\varphi = \int_{[0,1]^2} f(u_1, u_2) du_1 du_2$$

$\Rightarrow$ If $\varphi = 0$ then $f = 0$ a.e. and G is empty. Sparse but boring

$\Rightarrow$ Else $\varphi > 0$ and (in expectation) $|\mathcal{E}| = \varphi \times \binom{N_v}{2} = \Omega(N_v^2)$

- Not great, but in practice main barrier is vertex exchangeability
  - Suitable for unlabeled graphs, yet in many graphs labels matter
  - Can incorcorporate vertex attributes as covariates [Sweet'15]

Estimating SBM parameters

- SBMs defined up to parameters $\{\alpha_q\}_{q=1}^Q$ and $\{\pi_{qr}\}_{1\leq q,r\leq Q}$
- Conceptually useful to think about two sets of 'observations'
  - $\Rightarrow$ Latent class labels: $\mathbf{Z} = \{\{Z_{iq}\}_{q=1}^Q\}_{i\in\mathcal{V}}$, where $Z_{iq} = \mathbb{I}\{i \in \mathcal{C}_q\}$
  - $\Rightarrow$ Relational ties: $\mathbf{A} = [A_{ij}]$, where $A_{ij} = \mathbb{I}\{(i,j) \in \mathcal{E}\}$
- **But we only observe A**, recall $\mathbf{Z}$ are latent. Q assumed given
  - $\Rightarrow$ Interest both in parameter estimation and in vertex clustering

Model-based community detection

Suppose G adheres to an SBM with Q classes. Predict class membership labels $\mathbf{Z} = \{\{Z_{iq}\}_{q=1}^Q\}_{i\in\mathcal{V}}$, given observations $\mathbf{A} = \mathbf{a}$.

Maximum likelihood estimation

- If we were to observe $\mathbf{A} = \mathbf{a}$ and $\mathbf{Z} = \mathbf{z}$, the log-likelihood would be

$$\ell(\mathbf{a}, \mathbf{z}; \boldsymbol{\theta}) = \sum_i \sum_q z_{iq} \log \alpha_q + \frac{1}{2} \sum_{i \neq j} \sum_{q \neq r} z_{iq} z_{jr} \log b(A_{ij}; \pi_{qr})$$

$\Rightarrow$ Defined $\boldsymbol{\theta} = \{\{\alpha_q\}, \{\pi_{qr}\}\}$ and $b(a; \pi) = \pi^a (1 - \pi)^{1-a}$

- But we do not. Instead have to work with the observed data likelihood

$$\ell(\mathbf{a}; \boldsymbol{\theta}) = \log \left(\sum_{\mathbf{z}} \exp \{\ell(\mathbf{a}, \mathbf{z}; \boldsymbol{\theta})\} \right)$$

$\Rightarrow$ Unfortunately, evaluation of $\ell(\mathbf{a}; \boldsymbol{\theta})$ is typically intractable

- Mixture model viewpoint suggests an E-M procedure [Snijders'97]

$\Rightarrow$ Alternate between estimation of $\mathbb{E}\left[Z_{iq} \,\middle|\, \mathbf{A} = \mathbf{a}\right]$ and $\boldsymbol{\theta}$

$\Rightarrow$ Does not scale beyond $Q = 2$, $\mathrm{P}\left(\mathbf{Z} \,\middle|\, \mathbf{A} = \mathbf{a}\right)$ expensive

Variational maximum likelihood

- Variational approach to optimize a lower bound of $\ell(\mathbf{a};\boldsymbol{\theta})$, namely

$$J(R_{\mathbf{a}};\boldsymbol{\theta}) = \ell(\mathbf{a};\boldsymbol{\theta}) - \mathrm{KL}(R_{\mathbf{a}}(\mathbf{Z}), \mathrm{P}\left(\mathbf{Z} \,\middle|\, \mathbf{A} = \mathbf{a}\right))$$

  - KL denotes de Kullback–Leibler divergence
  - $R_{\mathbf{a}}(\mathbf{Z})$ is a tractable approximation of $\mathrm{P}\left(\mathbf{Z} \,\middle|\, \mathbf{A} = \mathbf{a}\right)$

- Mean field approximation to the conditional distribution

$$R_{\mathbf{a}}(\mathbf{Z}) = \prod_{i=1}^{N_v} h(\mathbf{Z}_i;\boldsymbol{\tau}_i)$$

  - $h(\cdot;\boldsymbol{\tau}_i)$: multinomial pmf with parameter $\boldsymbol{\tau}_i = [\tau_{i1}, \ldots, \tau_{iQ}]^\top$

J. J. Daudin et al, "A mixture model for random graphs," *Stat. Comput.*, vol. 18, 2008

Alternating maximization algorithm

Proposition: Given $\boldsymbol{\theta}$, the optimal variational parameters $\{\hat{\boldsymbol{\tau}}_i\} = \text{argmax}_{\{\boldsymbol{\tau}_i\}} J(R_{\mathbf{a}}; \{\boldsymbol{\tau}_i\}, \boldsymbol{\theta})$ satisfy the following fixed-point relation

$$\hat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_r b(A_{ij}; \pi_{qr})^{\hat{\tau}_{jr}}$$

Given $\{\boldsymbol{\tau}_i\}$, the values of $\boldsymbol{\theta}$ that maximize $J(R_{\mathbf{a}}; \{\boldsymbol{\tau}_i\}, \boldsymbol{\theta})$ are

$$\hat{\alpha}_q = \frac{1}{N_v} \sum_i \hat{\tau}_{iq}, \quad \hat{\pi}_{qr} = \sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jr} A_{ij} \Big/ \sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jr}$$

- Algorithm alternates between updates of $\boldsymbol{\theta}$ and $\{\boldsymbol{\tau}_i\}$ as follows

$$\boldsymbol{\theta}[k+1] = \underset{\boldsymbol{\theta}}{\text{argmax}}\, J(R_{\mathbf{a}}; \{\boldsymbol{\tau}_i[k]\}, \boldsymbol{\theta})$$
$$\{\boldsymbol{\tau}_i[k+1]\} = \underset{\{\boldsymbol{\tau}_i\}}{\text{argmax}}\, J(R_{\mathbf{a}}; \{\boldsymbol{\tau}_i\}, \boldsymbol{\theta}[k+1])$$

  - The sequence of J values is non-decreasing [Daudin et al'08]
  - Consistency results available as $N_v \to \infty$, Q fixed [Celisse et al'12]

Choice of the number of classes

- Number of classes Q often unknown and should be estimated
  - $\Rightarrow$ Use principles of Bayesian model selection
  - $\Rightarrow$ Prior $g(\boldsymbol{\theta} \,|\, m_Q)$ on $\boldsymbol{\theta}$ given the SBM m_Q has Q classess
- Integrated Classification Likelihood (ICL) criterion yields

$$\begin{aligned}\mathrm{ICL}(m_Q) = & \max_{\boldsymbol{\theta}} \log \mathcal{L}(\mathbf{a}, \hat{\mathbf{z}}(\boldsymbol{\theta}) \,|\, \boldsymbol{\theta}, m_Q) \\ & - \frac{Q(Q+1)}{4} \log \frac{N_v(N_v - 1)}{2} - \frac{Q-1}{2} \log N_v\end{aligned}$$

- Asymptotic approximation of the complete-data integrated likelihood

$$\mathcal{L}(\mathbf{a}, \mathbf{z} \,|\, m_Q) = \int \mathcal{L}(\mathbf{a}, \mathbf{z} \,|\, \boldsymbol{\theta}, m_Q) g(\boldsymbol{\theta} \,|\, m_Q) d\boldsymbol{\theta}$$

Graphon estimation

- **Goal:** estimate graphon f from observed realization G^{obs}
- Non-parametric regression approaches estimate f given $\{A_{ij}, U_i, U_j\}_{i,j \in \mathcal{V}}$
  $\Rightarrow$ Challenge is that the design points $U_1, \ldots, U_{N_v}$ are latent

SBM approximation

C. Gao et al, "Rate-optimal graphon estimation," *Annals of Statistics,* vol. 43, 2015

Histogram estimator (ordering and smoothing)

S. H. Chan and E. M. Airoldi, "A consistent histogram estimator for exchangeable graph models," *ICML*, 2014

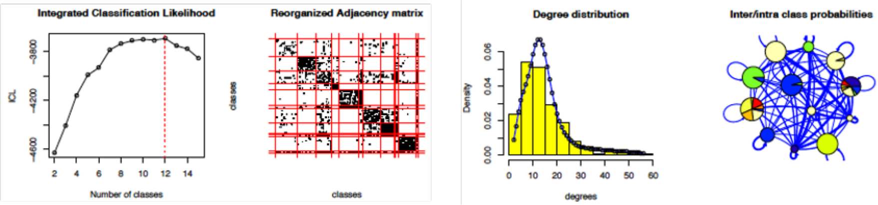
Gaussian process model

P. Orbanz and D. M. Roy, "Bayesian models of graphs, arrays and other exchangeable random structures," *IEEE Trans. PAMI,* vol. 37, 2015

Assessing goodness-of-fit

- Goodness-of-fit diagnostics $\Rightarrow$ mostly computational, visualization based
- Ex: French political blog network from October 2006 [Kolaczyk'17]

  $\Rightarrow$ We fit an SBM using variational MLE (**mixer** in R)



- Optimal value $\hat{Q} = 12$, but $Q \in [8, 12]$ reasonable (9 political parties)

  $\Rightarrow$ Permuted adjacency shows group structure (room for merging few)
- Relatively good fit of the degree distribution

Extensions of SBMs

Degree-corrected SBMs

- Communities with broad degree distributions

B. Karrer B and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E.*, vol. 83, 2011

Mixed-membership SBMs

- Nodes may belong only partially to more than one class

E. M. Airoldi, "Mixed membership stochastic blockmodels," *J. Machine Learning Research*, vol. 9, 2008

Hierarchical SBMs

- Hierarchical clustering meets SBMs

A. Clauset et al, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, 2008

Roadmap

Random graph models

Small-world models

Network-growth models

Exponential random graph models

Latent network models

Random dot product graphs

Random dot product graphs

- Consider a latent space $\mathcal{X}_d \subset \mathbb{R}^d$ such that for all

$$\mathbf{x}, \mathbf{y} \in \mathcal{X}_d \quad \Rightarrow \quad \mathbf{x}^\top \mathbf{y} \in [0, 1]$$

$\Rightarrow$ Inner-product distribution $F : \mathcal{X}_d \mapsto [0, 1]$

- Random dot product graphs (RDPGs) are defined as follows:

$$\mathbf{x}_1, \ldots, \mathbf{x}_{N_v} \overset{\text{i.i.d.}}{\sim} F,$$
$$A_{ij} \,\big|\, \mathbf{x}_i, \mathbf{x}_j \sim \text{Bernoulli}(\mathbf{x}_i^\top \mathbf{x}_j)$$

for $1 \leq i, j \leq N_v$, where $A_{ij} = A_{ji}$ and $A_{ii} \equiv 0$

- A particularly tractable latent position random graph model

$\Rightarrow$ Vertex positions $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}]^T \in \mathbb{R}^{N_v \times d}$

S. J. Young and E. R. Scheinerman, "Random dot product graph models for social networks," *WAW*, 2007

Connections to other models

- RDPGs ecompass several other classic models for network graphs

  Ex: Recover Erdös-Renyi $G_{N_v,p}$ graphs with $d = 1$ and $\mathcal{X}_d = \{\sqrt{p}\}$

  Ex: Recover SBM random graphs by constructing F with pmf

$$\mathrm{P}\left(\mathbf{X} = \mathbf{x}_q\right) = \alpha_q, \quad q = 1, \ldots, Q$$

  after selecting d and $\mathbf{x}_1, \ldots, \mathbf{x}_Q$ such that $\pi_{qr} = \mathbf{x}_q^\top \mathbf{x}_r$

- Approximation results for SBMs justify the expressiveness of RDPGs
- RDPGs are special cases of latent position models [Hoff et al'02]

$$A_{ij} \,\big|\, \mathbf{x}_i, \mathbf{x}_j \sim \text{Bernoulli}(\kappa(\mathbf{x}_i, \mathbf{x}_j))$$

  $\Rightarrow$ Approximate these accurately for large enough d [Tang et al'13]

Estimation of latent positions

- **Q:** Given G from an RDPG, find the 'best' $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}]^\top$?
- MLE is well motivated but it is intractable for large N_v

$$\hat{\mathbf{X}}_{ML} = \underset{\mathbf{X}}{\operatorname{argmax}} \prod_{i<j} (\mathbf{x}_i^\top \mathbf{x}_j)^{A_{ij}} (1 - \mathbf{x}_i^\top \mathbf{x}_j)^{1-A_{ij}}$$

- Instead, let $P_{ij} = \mathrm{P}\left((i,j) \in \mathcal{E}\right)$ and define $\mathbf{P} = [P_{ij}] \in [0,1]^{N_v \times N_v}$
  - $\Rightarrow$ The RDPG model specifies that $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$
  - $\Rightarrow$ **Key:** Observed $\mathbf{A}$ is a noisy realization of $\mathbf{P}$ $(\mathbb{E}[A] = \mathbf{P})$
- Suggests a LS regression approach to find $\mathbf{X}$ s.t. $\mathbf{X}\mathbf{X}^\top \approx \mathbf{A}$

$$\hat{\mathbf{X}}_{LS} = \underset{\mathbf{X}}{\operatorname{argmin}} \|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_F^2$$

Adjacency spectral embedding

- Since $\mathbf{A}$ is real and symmetric, can decompose it as $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$
  - $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_{N_v}]$ is the orthogonal matrix of eigenvectors
  - $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_{N_v})$, with eigvenvalues $\lambda_1 \geq \ldots \geq \lambda_{N_v}$
- Define $\hat{\boldsymbol{\Lambda}} = \text{diag}(\lambda_1^+, \ldots, \lambda_d^+)$ and $\hat{\mathbf{U}} = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$ $(\lambda^+ := \max(0, \lambda))$
- Best rank-d, positive semi-definite (PSD) approximation of $\mathbf{A}$ is $\hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}}^\top$

  $\Rightarrow$ Ajacency spectral embedding (ASE) is $\hat{\mathbf{X}}_{LS} = \hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}^{1/2}$ since

$$\mathbf{A} \approx \hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}}^\top = \hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}^{1/2}\hat{\boldsymbol{\Lambda}}^{1/2}\hat{\mathbf{U}}^\top = \hat{\mathbf{X}}_{LS}\hat{\mathbf{X}}_{LS}^\top$$

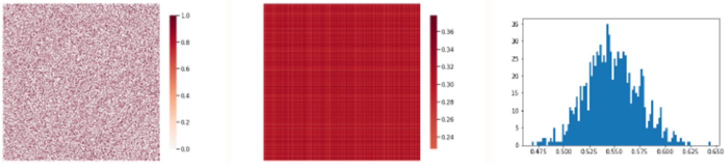
- **Q:** Is the solution unique? Nope, inner-products are rotation invariant

$$\mathbf{P} = \mathbf{X}\mathbf{W}(\mathbf{X}\mathbf{W})^\top = \mathbf{X}\mathbf{X}^\top, \quad \mathbf{W}\mathbf{W}^\top = \mathbf{I}_d$$

  $\Rightarrow$ RDPG embedding problem is identifiable modulo rotations

Embedding an Erdös-Renyi graph

- Ex: Erdös-Renyi graph $G_{1000,0.3}$, realization of $\mathbf{A}$ (left)



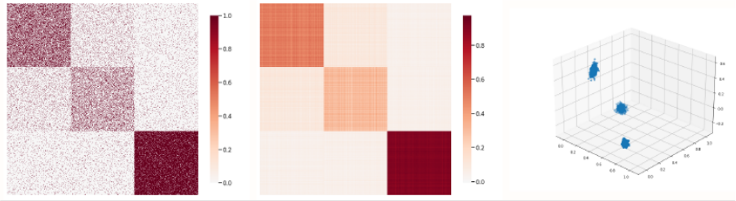
- For $d = 1$ we compute the ASE $\hat{\mathbf{x}}_{LS}$ and plot $\hat{\mathbf{x}}_{LS}\hat{\mathbf{x}}_{LS}^\top$ (center)
  - $\Rightarrow$ Appoximates well the constant matrix $\mathbf{P} = 0.3 \times \mathbf{1}\mathbf{1}^\top$
  - $\Rightarrow$ Suported by histogram of entries in $\hat{\mathbf{x}}_{LS}$ (right, $\sqrt{p} = 0.547$)
- Consisentcy and limiting distribution results for ASEs available

A. Athreya et al., "Statistical inference on random dot product graphs: A survey," *J. Mach. Learn. Res.*, vol. 18, pp. 1-92, 2018

Embedding an SBM graph

- Ex: SBM with $N_v = 1500$, $Q = 3$ and mixing parameters

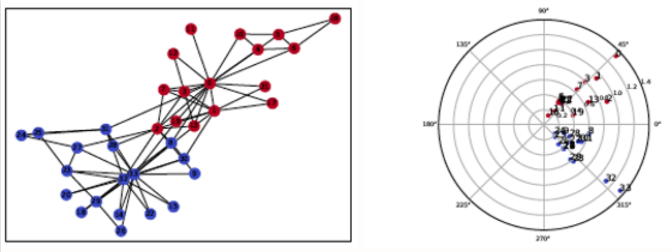
$$\boldsymbol{\alpha} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \quad \boldsymbol{\Pi} = \begin{bmatrix} 0.5 & 0.1 & 0.05 \\ 0.1 & 0.3 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}$$



- Sample adjacency (left), $\hat{\mathbf{X}}_{LS}\hat{\mathbf{X}}_{LS}^\top$ (center), rows of $\hat{\mathbf{X}}_{LS}$ (right)
- Use embeddings to bring to bear geometric methods of analysis

Interpretability of the embeddings

- Ex: Zachary's karate club graph with $N_v = 34$, $N_e = 78$ (left)

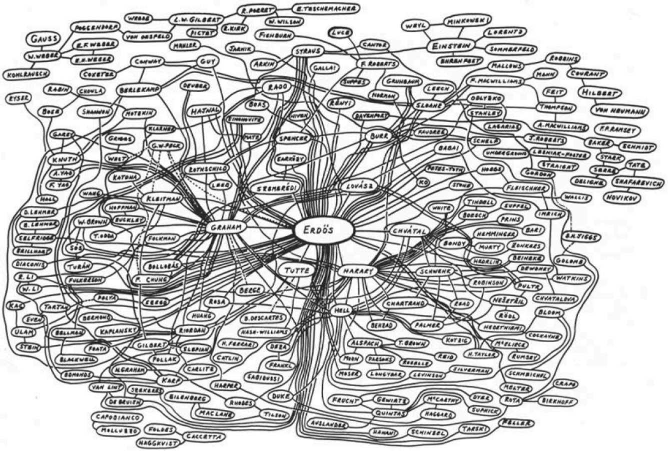


- Node embeddings (rows of $\hat{\mathbf{X}}_{LS}$) for $d = 2$ (right)
  - Club's administrator ($i = 0$) and instructor ($j = 33$) are orthogonal
- Interpretability of embeddings a valuable asset for RDPGs

  $\Rightarrow$ Vector magnitudes indicate how well connected nodes are

  $\Rightarrow$ Vector angles indicate positions in latent space

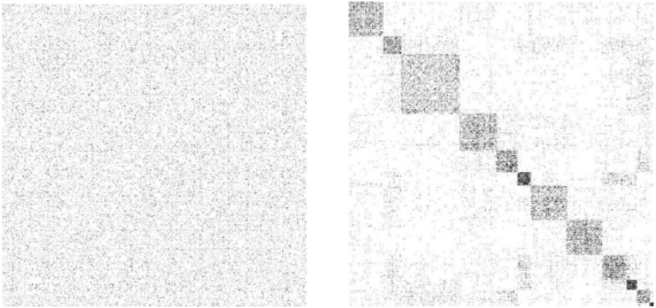
- Ex: Mathematics collaboration network centered at Paul Erdös



- Most mathematicians have an Erdös number of at most 4 or 5
  - ⇒ Drawing created by R. Graham in 1979

Mathematicians collaboration graph

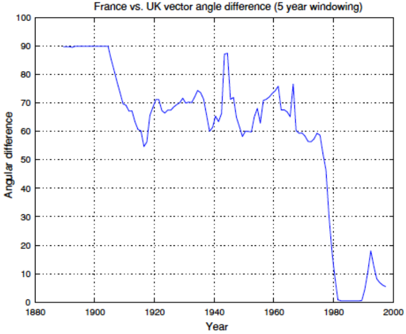
- Coauthorship graph G, $N_v = 4301$ nodes with Erdös number ≤ 2
  - $\Rightarrow$ No discernible structure from the adjacency matrix $\mathbf{A}$ (left)



- Community structure revealed after row-colum permutation (right)
  - (i) Obtained the ASE $\hat{\mathbf{X}}_{LS}$ for the mathematicians
  - (ii) Performed angular k-means on $\hat{\mathbf{X}}_{LS}$'s rows [Scheinerman-Tucker'10]

International relations

- Ex: Dynamic network G_t of international relations among nations
  $\Rightarrow$ Nations $(i, j) \in \mathcal{E}_t$ if they have an alliance treaty during year t



- Track the angle between UK and France's ASE from 1890–1995
  - Orthogonal during the late 19th century
  - Came closer during the wars, retreat during Nazi invasion in WWII
  - Strong alignment starts in the 1970s in the run up to the EU

Closing remarks and extensions

- Neglected the constraint $[\hat{\mathbf{X}}_{LS}\hat{\mathbf{X}}_{LS}^\top]_{ii} = 0$. Fix via iterative algorithm

  E. R. Scheinerman and K. Tucker, "Modeling graphs using dot product representations," *Comput. Stat.*, vol. 25, pp. 1-16, 2010

- Assumed $\mathbf{A}$ to be PSD. Extension known as generalized RDPG

  P. Rubin-Delanchy et al, "A statistical interpretation of spectral embedding: The generalised random dot product graph," *arXiv:1709.05506 [stat.ML]*, 2017

- RDPG variants to model weighted and directed graphs possible

  F. Larroca et al, "Change point detection in weighted and directed random dot product graphs," *ICASSP*, 2021

- Host of applications in testing, clustering, change-point detection, ...

Glossary

- Network graph model
- Random graph models
- Configuration model
- Matching algorithm
- Switching algorithm
- Model-based estimation
- Assessing significance
- Reference distribution
- Network motif
- Small-world network
- Decentralized search
- Watts-Strogatz model
- Time-evolving network
- Network-growth models
- Preferential attachment
- Barabási-Albert model
- Copying models
- Exponential family
- Exponential random graph models
- Configurations
- Network statistic
- Homogeneity
- Markov random graphs