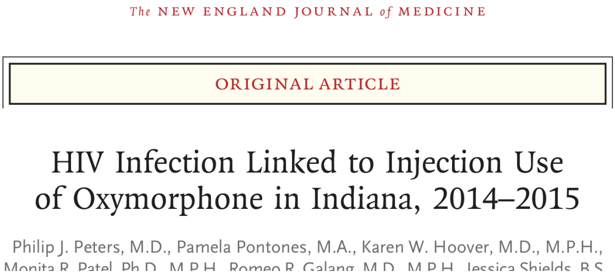
Network models

Forrest W. Crawford

Why Networks?



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

HIV Infection Linked to Injection Use of Oxymorphone in Indiana, 2014–2015

Philip J. Peters, M.D., Pamela Pontones, M.A., Karen W. Hoover, M.D., M.P.H.,
Monita R. Patel, Ph.D., M.P.H., Romeo R. Galang, M.D., M.P.H., Jessica Shields, B.S.

Figure 1

Networks

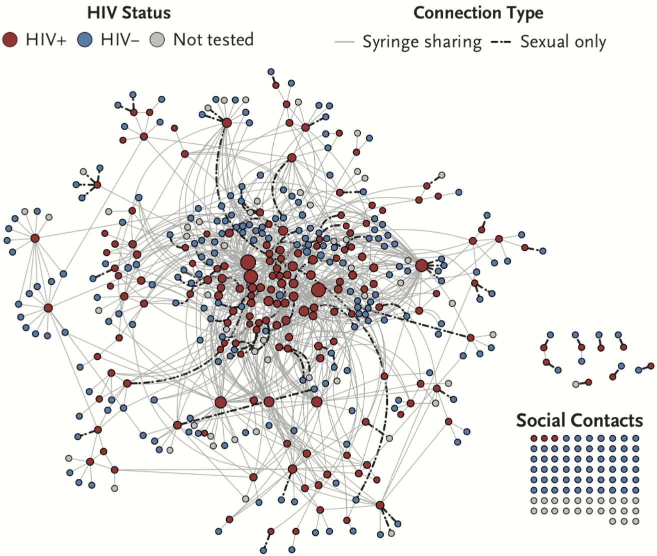


Figure 2

Networks

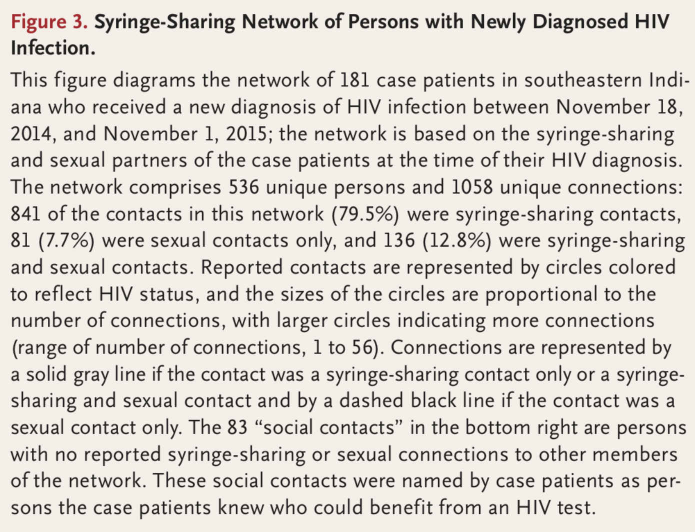


Figure 3. Syringe-Sharing Network of Persons with Newly Diagnosed HIV Infection.

This figure diagrams the network of 181 case patients in southeastern Indiana who received a new diagnosis of HIV infection between November 18, 2014, and November 1, 2015; the network is based on the syringe-sharing and sexual partners of the case patients at the time of their HIV diagnosis. The network comprises 536 unique persons and 1058 unique connections: 841 of the contacts in this network (79.5%) were syringe-sharing contacts, 81 (7.7%) were sexual contacts only, and 136 (12.8%) were syringe-sharing and sexual contacts. Reported contacts are represented by circles colored to reflect HIV status, and the sizes of the circles are proportional to the number of connections, with larger circles indicating more connections (range of number of connections, 1 to 56). Connections are represented by a solid gray line if the contact was a syringe-sharing contact only or a syringe-sharing and sexual contact and by a dashed black line if the contact was a sexual contact only. The 83 “social contacts” in the bottom right are persons with no reported syringe-sharing or sexual connections to other members of the network. These social contacts were named by case patients as persons the case patients knew who could benefit from an HIV test.

Figure 3

Networks

Sometimes, instead of data about isolated units (humans, usually) and their attributes/outcomes, we have relational data about the connections between units.

Networks are *data structures* for representing relational information.

Some people think networks are transcendental data structures that can reveal many hidden truths about the human condition. I don't, but I do think networks are very useful.

But there are a lot of pitfalls in analyzing network data. Many data analysis heuristics and existing statistical tools do not work for network data, and may give misleading results.

Network notation

Mathematicians call networks *graphs*. A graph consists of: - a set V of *vertices* (nodes, subjects, units, egos) - a set E of *edges* (links, arcs, connections, relationships)

$$E = \{\{i, j\} \text{ for } i, j \in V\}$$

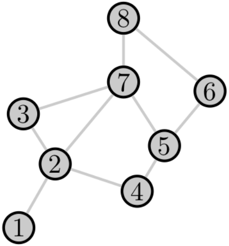


Figure 4

$$V = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

Examples of networks $G = (V, E)$

Network (G) & Vertices (V) & Edges (E)

Social	people	friendships
Needle	needles	people
Publication	articles	citations
Publication	authors	co-authorship
Phylogenetic	species	ancestral relationship
Trade	countries	trade relationships
Gene expression	genes	interactions
Brain	cortical regions	fiber tracts
Neural	neurons	axons
Injection	people	needle sharing partners

Types of networks: directed/undirected

An undirected network has symmetric links

Types of networks: weighted

In a *weighted* network, each edge $\{i, j\} \in E$ has a numeric attribute z_{ij}.

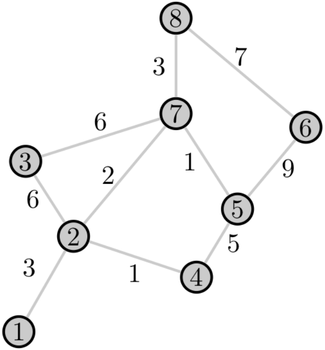


Figure 7

Sometimes we define $z_{ij} = 0$ if $\{i, j\} \notin E$.

Special types of graphs

Graphs can be

- empty ($E = \emptyset$)
- complete ($E = \{\{i, j\}$ for all $i, j \in V\}$)
- connected (all vertices are connected to each other via a path)
- acyclic (tree, DAG)
- planar (no crossing edges)
- sparse (not too many edges)
- dense (lots of edges)

There are lots of other types of graphs!

Terminology: degree

The *degree* of a vertex is the number of edges incident to it:

$$d_i = |\{j \in V : \{i,j\} \in E\}|$$

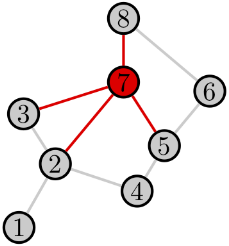


Figure 8

Vertex 7 has degree $d_7 = 4$. Sometimes social scientists call this the *egocentric network size* or just *network size*.

For directed graphs, we distinguish between

- *in-degree* $d^{\text{in}} = |\{j : (j, i) \in E\}|$

Degree sequence/distribution

The *degree sequence* of a graph is the sequence (sometimes ordered) of vertex degrees.

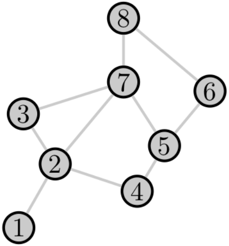


Figure 9

Degree sequence: $\mathbf{d} = (1, 4, 2, 2, 3, 2, 4, 2)$

Does the degree sequence $\mathbf{d}$ uniquely determine the topology of G?

The *degree distribution* is the frequency of each degree value in a graph.

More terminology

- A network *motif* is just a pattern of edges for a collection of vertices.
- A *clique* is a set of vertices for which the edge set is *complete*.
- The *giant component* is the biggest connected component of a graph.
- *Transitivity* is the propensity of the graph to exhibit *triangles*.
- A network with *community structure* has more links within groups of vertices than between groups.
- *Small world* networks have every pair of vertices connected by a short path.
- The graph *Laplacian* is $L = D - A$ where $D = \mathrm{diag}(d_1, \ldots, d_n)$ and A is the adjacency matrix.
- The *eigenvalues* of A or L can tell you about how *connected* the graph is.

Representing networks

Consider a graph $G = (V, E)$.

Adjacency matrix: $A_{ij} = 1$ if and only if $\{i, j\} \in E$.

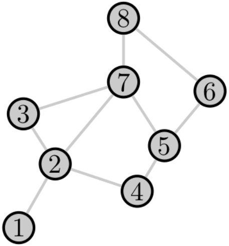
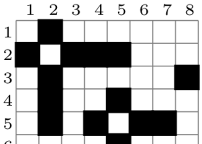


Figure 10



	1	2	3	4	5	6	7	8
1		■						
2	■		■	■	■			
3		■						■
4		■			■			
5		■		■		■	■	
6					■			

Properties of the adjacency matrix

In general, using the adjacency matrix representation of a graph lets you do *linear algebra*.

Simple properties:

- $A1 = \mathbf{d}$
- $1'A1 = 2|E|$ (handshake theorem)
- The i, j element of A^k is the number of walks of length k that start at i and end at j. What is trace$(A^3)/6$?

Spectral properties

- Recall that λ is an eigenvalue and $x \neq 0$ is an eigenvector if $Ax = \lambda x$
- When A is symmetric, it has $|V|$ real eigenvalues, called the “spectrum” of G.
- Certain useful properties, like connectedness and clustering, can be deduced from the spectrum of G.

Viewing networks: principles

Visualizing a network requires a *layout algorithm* that tells your computer where to put each of the vertices relative to each other in two dimensions.

Layout algorithms try to place vertices in 2 dimensions so that:

- vertices do not overlap
- few edges overlap
- network components are evident
- community structure is evident
- path length corresponds to spatial distance But, it is *very* difficult to achieve all these goals simultaneously.

I recommend:

- Use visualization to gain insight, not to make arguments.
- Start with a circular-by-degree layout.

One network, many hairballs

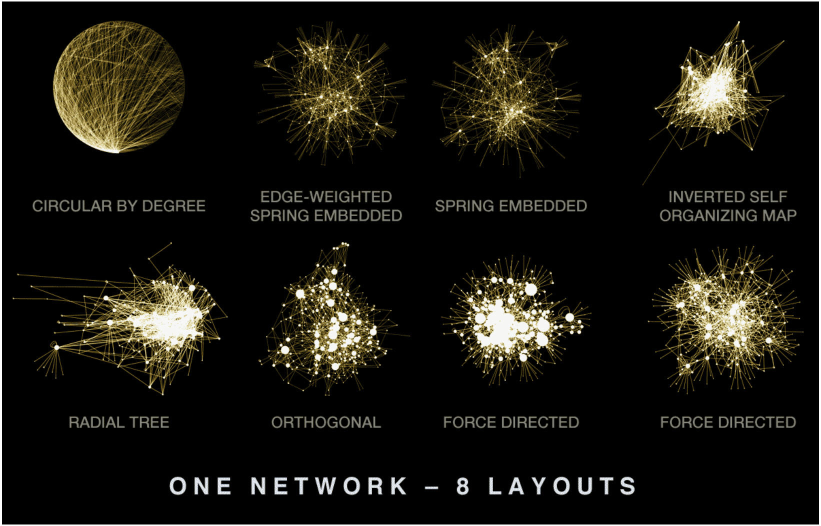


Figure 12

Viewing networks: algorithms and pitfalls

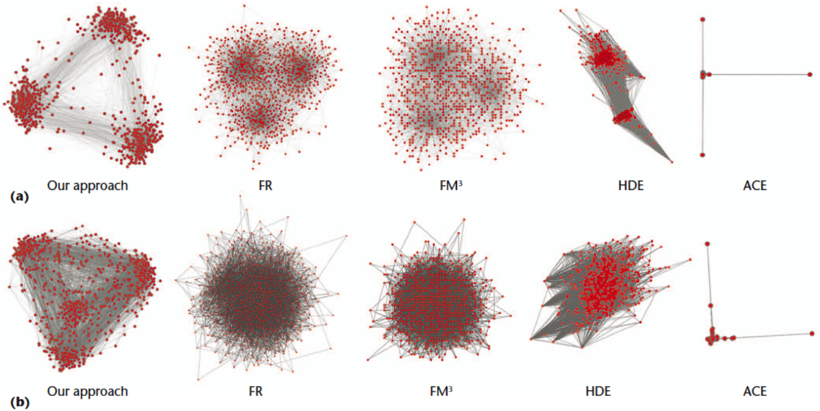


Figure 13

Hu, Lu, Wu, "Spectrum-Based Network Visualization for Topology Analysis", *IEEE Computer Graphics and Applications* 33, pages 58-68, 2013.

Viewing weighted networks

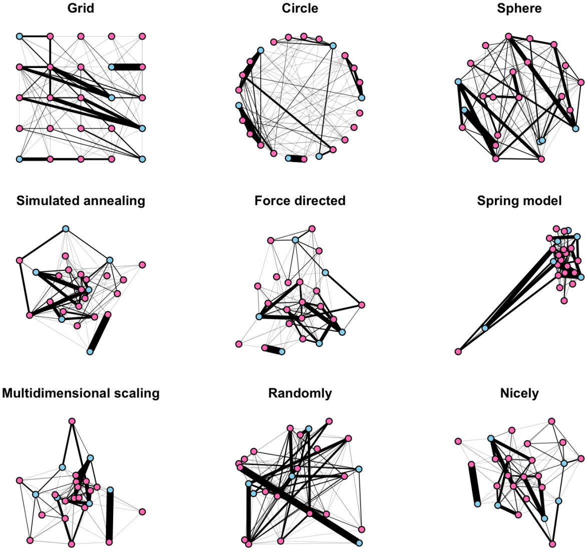


Figure 14

Network summaries: centrality

For a vertex i, define - *degree centrality*: d_i - *closeness centrality*:

$$\frac{1}{\sum_{j \in V} \text{shortestPathDist}(i, j)}$$

- *betweenness centrality*:

$$\sum_{j \neq k \neq i} \frac{\#\text{shortestPaths}(j \to i \to k)}{\#\text{shortestPaths}(j \to k)}$$

- *eigenvector centrality*: x_i where

$$Ax = \lambda x$$

defines the eigenvector x and its corresponding (largest) eigenvalue λ. There are many other measures of centrality.

Network summaries: clustering/community detection

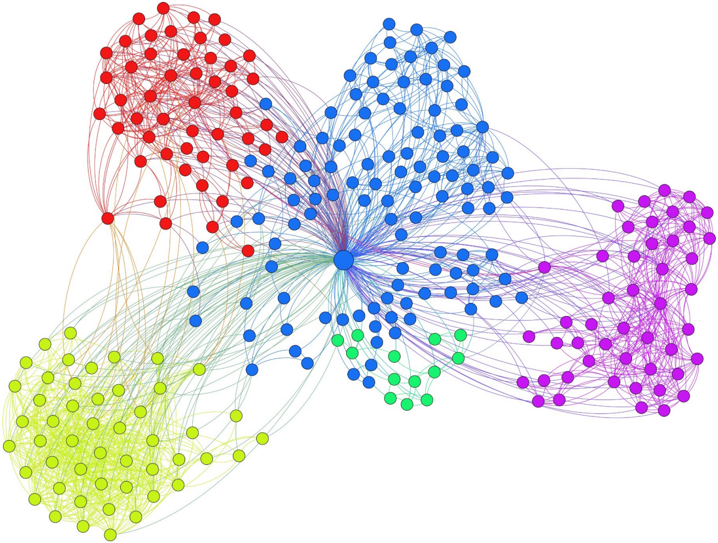


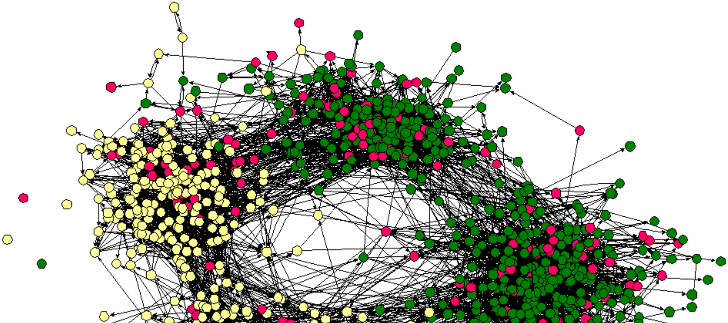
Figure 15

Vertex and edge attributes

Vertices and edges can have *attributes*, or observations associated with them.

- Let x_i be attributes of vertex $i \in V$. - Let z_{ij} be attributes of the edge $\{i, j\}$.

We might want to summarize node traits in the network by - degree assortativity: connections to nodes with similar degree - trait assortativity (homophily): connections to nodes with similar traits



Network models

Why do we need models for networks? Networks are complicated, and there are lots of them. By constructing models, learn about the structure of a complicated object using a small number of parameters.

How many networks?

One reason we need models is that there are *a lot* of networks:

- Number of undirected graphs on n vertices: $2^{\binom{n}{2}}$
- Number of directed graphs on n vertices: $2^{n(n-1)}$

Example For a graph $G = (V, E)$ with $n = |V| = 50$, there are approximately 5.77×10^{368} possible undirected graphs and 3.34×10^{737} directed graphs. There are about 10^{80} atoms in the known universe.

This means that it can be prohibitively difficult to enumerate them for even modest n.

So it is helpful to have a model with fewer parameters than $2^{\binom{n}{2}}$ that helps us summarize networks.

Network models

A *network model* is a set of networks $\mathcal{G}$ and a probability function $\Pr(\cdot) \geq 0$ that maps networks to probabilities such that

$$\sum_{G \in \mathcal{G}} \Pr(G) = 1.$$

Probabilistic network models and parameter estimation: If you have a probability model $\Pr(\cdot|\theta)$, where θ is a set of parameters, and you observe a network G, you might want to find the value of θ that maximizes the probability of the network. In other words,

$$\hat{\theta} = \arg\max_{\theta} \Pr(G|\theta)$$

This is called “maximum likelihood”.

There are many other ways to estimate parameters. %But the basic idea is to use an observed network (G, X, Z) to tell you something about the model $\Pr(G, X, Z|\theta)$ that might have generated that network.

Network model: Erdos-Renyi

Consider a graph of $n = |V|$ vertices in which each edge exists independently with probability p. Then the probability of observing a particular graph $G = (V, E)$ is

$$\Pr(G|p) = p^{|E|}(1-p)^{\binom{n}{2}-|E|}$$

Many analytic results can be derived for the ER model:

- $|E| \sim \text{Binomial}\left(\binom{n}{2}, p\right)$
- $d_i \sim \text{Binomial}\left(n-1, p\right)$
- Expected degree of a vertex is $\mathrm{E}[d_i] = (n-1)p$
- Expected number of edges $\mathrm{E}[|E|] = \binom{n}{2}p$
- $p = 1/2$ gives the *uniform* distribution over graphs,

  $$\Pr(G|p = 1/2) = 2^{-\binom{n}{2}}$$

  Does $p = 1/2$ give the uniform distribution over $|E|$?

Also known as: Bernoulli model, homogeneous random graph

Network model: logistic edge probabilities

A simple model for graph edges, given vertex attributes, is

$$\log \left[\frac{\Pr(i \sim j)}{\Pr(i \nsim j)} \right] = \alpha + \beta f(x_i, x_j)$$

where α is an intercept and $f(x_i, x_j)$ is some (symmetric) function:

- $f(x_i, x_j) = x_i + x_j$
- $f(x_i, x_j) = ||x_i - x_j||$
- $f(x_i, x_j) = (\epsilon + ||x_i - x_j||)^{-1}$

You should use this model to explain the formation of a network if:

- the network was formed *after* vertex attributes arose
- homophily (or heterophily, or some kind of assortativity) is the only reason that links exist
- edges are conditionally independent given vertex attributes (no global structure)

Exponential random graph models (ERGMs)

An *exponential random graph model* (ERGM) has probability

$$\Pr(G|\theta) = \frac{\exp[\theta' h(G)]}{\kappa(\theta)}$$

where $h(G)$ is some function of G (called the *sufficient statistic*), and

$$\kappa(\theta) = \sum_{g \in \mathcal{G}} \exp[\theta' h(g)]$$

is a normalizing constant that ensures these probabilities sum to 1.

Good things:

- easy to parameterize in low dimension
- can capture complex global topological properties Bad things:
- intractable normalizing constant $\kappa(\theta)$
- difficulty fitting
- degeneracy

ERGMs

Examples of ERGM sufficient statistics

- $h(G) = |E|$ (Erdos-Renyi)
- $h(G) = (|E|, \#\text{triangles }\,)$ (dyads and triangles)
- $h(G) = (AKS_\lambda, \#\text{triangles})$ (alternating k-star statistic) with

$$AKS_\lambda(G) = \sum_{k=1}^{n-1} (-1)^k \frac{S_k(G)}{\lambda^{k-2}}$$

where $S_k(G)$ is the number of k-stars in G. It is *not* true that every function $h(G)$ yields a valid joint distribution over graphs $\mathcal{G}$.

ERGMs: fitting

We observe G, we want to fit the ERGM $\Pr(G|\theta)$ to learn about θ.

$$\ell(\theta) = \log\left[\Pr(G|\theta)\right] = \theta' h(G) - \log \kappa(\theta)$$

Taking the derivative (gradient) with respect to θ (and then doing some algebra) yields the estimating equation

$$\nabla_\theta \ell(\theta) = h(G) - \mathrm{E}_\theta[h(G)] = 0$$

This suggests a stochastic algorithm for learning about θ:

- For a given θ, simulate lots of graphs $G_k \sim \Pr(G|\theta)$:

$$\frac{1}{M}\sum_{k=1}^{M} h(G_k) \rightarrow \mathrm{E}_\theta[h(G)]$$

- Compute $h(G) - \mathrm{E}_\theta[h(G)]$.
- Choose another value of θ to make this difference smaller.
- Repeat

ERGMs: degeneracy

Model degeneracy occurs:

- when a combination of parameters θ implies that only a small number of graphs have substantial non-zero probabilities; often, these graphs are radically different from each other, for example the empty graph and the fully connected graph;
- when the MLE of θ does not existent or is hard to obtain, or the MLE of θ as computed by MCMC methods fails to converge or appears to converge extremely slowly;
- when the estimate of θ would make the observed network very unlikely.

Rinaldo, Fienberg, Zhou. On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics* 3 (2009): 446-484.

ERGMs: degeneracy example

Let $h(G) = (\#\text{edges}, \#\text{triangles})$. Let $\Pr(G|\theta) \propto \exp[\theta' h(G)]$.

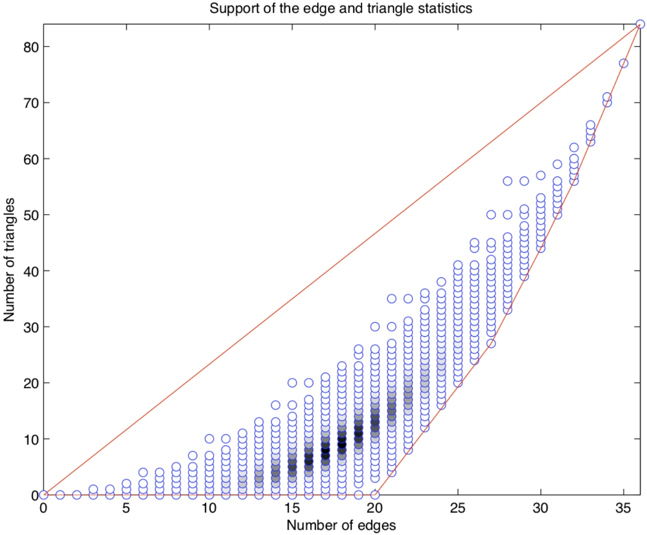
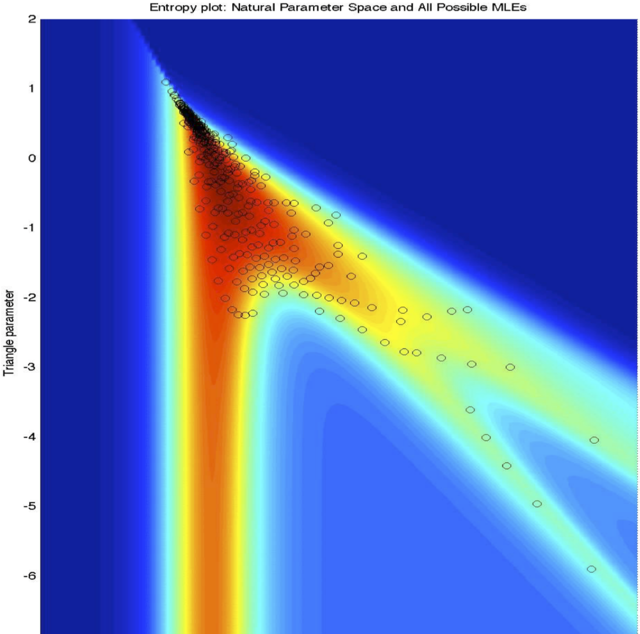


Figure 18

ERGM degeneracy



Exchangeable: SBM

Exchangeable graphs have the property that their distribution is invariant to permutation of the vertex labels.

The stochastic block model (SBM) is a graph in which each vertex $i \in V$ has a categorical attribute $x_i \in \{1, \ldots, K\}$. The attribute x_i is called the "block" membership of i.

Define a $K \times K$ matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{pmatrix}$$

that defines the edge probabilities between each of the K "blocks".

An edge between i and j is formed with probability P_{x_i, x_j}.

Exchangeable: SBM

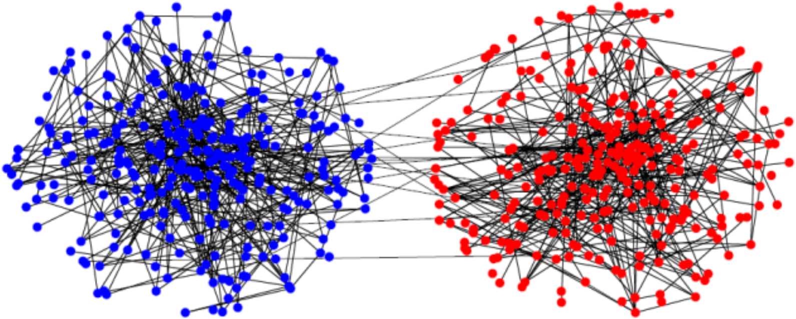


Figure 20

See also: degree-corrected SBM.

Other network models

There are lots of other network models designed to capture important features of real-world networks.

One of the main features is a “heavy-tailed” degree distribution in which most vertices have small degree, and some have very large degree.

Further reading:

- Scale-free, power-law, Barab’asi-Albert, preferential attachment
- Small world, Watts-Strogatz
- Exchangeable (graphon)
- Degree-corrected SBM

Models for processes on networks

Dynamic or stochastic processes on networks are beyond the scope of this talk. However, you might be interested in further reading on:

- Random walks on networks, and their equilibrium distributions
- Markov random fields as coherent models for vertex outcomes
- Clifford-Hammersley theorem, Brooks' Lemma for existence of valid joint distributions on vertex outcomes
- Diffusions, cascades
- Epidemics on networks
- (Economic) Games on networks, equilibria

Analysis: first steps

First, think about the data. For example:

- How was the network collected?
- Are any nodes/edges missing/unobserved?
- Are any node/edge attributes missing?
- How did the network form?
- Is the network static in time, or dynamic?
- How did the attributes arise: all at once, or over time?

Next,

- Compute some summaries: attribute means, standard deviations, histograms
- Compute assortativity with respect to degree and other attributes
- Compute topological summaries: degree distribution, density, triangles
- Visualize the network in several different ways

Analysis pitfalls: network sub-samples

Beware the sub-sampled network!

Suppose $G = (V, E)$ is a graph from a graph model $\Pr(G)$. Suppose you obtain a randomly chosen vertex subset $V_S \subset V$. You observe the _induced subgraph} $G_S = (V_S, E_S)$, where $\{i, j\} \in E_S$ if and only if $i, j \in V_S$ and $\{i, j\} \in E$.

Then in general G_S does _not} follow the graph model $\Pr(G)$.

So, *you cannot generalize findings from G_S to G*. In particular, sub-samples destroy triangles and other higher-order structure, so inferences about the topology of G_S are not representative of the topology of G.

Exception: exchangeable models like ER and SBM.

See also: Chapter 5 of Kolaczyk, Statistical Analysis of Network Data.

Analysis pitfalls: causation, homophily, contagion

Suppose vertex attributes have an effect on both edge existence and vertex outcome. *Then conditioning on the network A induces non-causal association between attributes and outcomes*.

This is also called “collider” or “selection”" bias. Be extra careful about ascribing a causal interpretation to “spillover” estimates in networks.

Auto-logistic model for binary outcomes

The standard network regression for a binary vertex outcome y_i is usually:

$$\log\left(\frac{\Pr(Y_i = 1)}{\Pr(Y_i = 0)}\right) = \alpha + x_i'\beta + \frac{\gamma}{|N_i|}\sum_{j \in N_i} y_j$$

where α (intercept); β (individual effect); N_i (neighbors of i); γ (peer/spillover effect)

You can use it if you agree with all of these statements:

- The outcomes y_i were realized simultaneously
- Yet, before realizing their outcome, i looked to see what their neighbors were up to
- i only cares about the proportion, not the number, of her neighbors who have positive outcome

Estimating $\gamma \neq 0$ does *not* imply contagion, spillover, or "peer effects".

Auto-normal linear model

Consider the linear model for the mean outcome

$$\mathsf{E}[Y_i] = \alpha + \beta' x_i + \frac{\gamma}{|N_i|} \sum_{j \in N_i} (y_j - \alpha - \beta' x_j)$$

with

$$\mathsf{Var}[Y_i] = \sigma^2$$

This defines a Normal (Gaussian) linear model. It produces a Markov random field and yields a valid joint distribution.

Estimating $\gamma \neq 0$ does *not* imply contagion, spillover, or "peer effects".

Network regression

General advice:

- Don't use vertex degree (or any topological feature) as a covariate in a regression, unless you believe the network does not induce dependence, and degree-exchangeability holds [Hint: it doesn't].
- Instead, use a measure of "network exposure" in regression models for vertex outcomes.
- If you must use alters' outcomes on the right-hand side of a regression equation, use the auto-normal or auto-logistic model. This results in a valid joint distribution (a Markov random field)
- If you invent your own conditional mean model $\mathrm{E}[Y_i]$ as a function of alters' Y_j's, use caution: you might be specifying a model for which no valid joint distribution exists.

Analysis pitfalls: bootstrap standard errors

When subjects are independent, researchers sometimes _resample} outcomes and compute the distribution of an estimator under resampling.

_Bootstrap standard errors} for independent units that do not depend on asymptotic normality can be computed in this way.

Suppose you have a statistic (estimator) $T(G)$.

Q: Can you resample vertices in G to get valid bootstrap standard errors? A: Basically never.

However, some classes of random graph model (exchangeable) have a distribution that is invariant to sampling subgraphs.

Principles of network data analysis

In no particular order:

- Sample size is probably $n = 1$, and definitely not $n = |V|$.
- Don't sub-sample networks, unless you're following Ch 5 of Kolaczyk
- Beware treating alters' outcomes as ego's predictors.
- Take the passage of time seriously in analyses of contagion.
- Report meaningful summaries.

Resources

books

- Eric D. Kolaczyk Statistical Analysis of Network Data
- Mark E. J. Newman Networks: An Introduction
- Matthew O. Jackson Social and Economic Networks

[Back: Agent-based Models] [Home] [Next: Network models]

References