Lecture 1: Conditionally-Independent Dyad Models

36-781, Fall 2016

25 October 2016

Abstract

These notes go over the same material as the actual lecture, with some amplifications, and a few corrections of stuff done from memory.

Contents

1 Scope of the course

This course is a fast-paced introduction to non-parametric statistical models of networks. It tries to bring people who are already familiar with network theory, with statistical modeling and with advanced probability reasonably close to the research frontier. We will *not* go back over such background material, except occasionally to

set the stage for new developments or to emphasize parts of the background which are usually neglected.

There are three leading mathematical ideas which have shaped the recent literature on non-parametric network models. These are the idea of network models in which dyads are conditionally independent given latent attributes of the nodes; the idea that any good network model should obey certain symmetry properties, and that all models with such symmetries can be decomposed into simple ("extremal") models with that symmetry, or into mixtures of them; and the notion of the limit of a sequence of graphs. All three ideas are intimately related, and much of the first part of the course will be about exploring the links between these notions.

This lecture is about the first idea, about conditionally-independent dyad models. We begin with the most basic, symmetric, independence-ful model possible, the homogeneous random graph; then we go to inhomogeneous random graphs, especially ones where the inhomogeneities are structured (and so allow for inference); finally, we consider graph models where all dyads are conditionally independent, and their symmetry properties.

Notation Throughout, n will refer to the number of nodes in a graph, and $\mathbf{A}$ will be the $n \times n$ random adjacency matrix, whose generic element is A_{ij}. Realizations of this random variable will be written as $\mathbf{a}$ (for the matrix) or a_{ij} (for the element).

For undirected graphs, (i, j) should be read as the un-ordered pair of nodes i and j (or **dyad**). For directed graphs, read (i, j) as the ordered pair. When S is a set of (i, j) pairs, A_S is the corresponding collection of edge-indicator variables.

2 Random graphs (reprise)

If you are reading these notes, you of course know the Erdos-Renyi model[1], which is defined by two properties:

1. All dyads are mutually independent: $A_{ij} \perp\!\!\!\perp A_{kl}$ unless $(i, j) = (k, l)$. More exactly, for any two sets of dyads S and T, $A_S \perp\!\!\!\perp A_T$ unless $S \cap T \neq \emptyset$.
2. All dyads are identically distributed: $\Pr\left(A_{ij} = 1\right) = p$.

This model lends itself to much genuinely beautiful mathematics, which have been studied exhaustively over the last sixty-odd years. It is also clearly amenable to statistical analysis. After all, $A_{ij} \sim_{IID} \text{Bernoulli}(p)$, so by the law of large numbers

$$\frac{1}{n^2} \sum_{(i,j)} A_{ij} \to p \tag{1}$$

[1] More properly, the Solomonoff-Rapoport-Erdos-Renyi-Gilbert model (Solomonoff and Rapoport, 1951; Erdös and Rényi, 1960; Gilbert, 1959).

with whatever mode of convergence you like[2]. Hence the maximum likelihood estimate of p is (strongly) consistent, unbiased, efficient, etc., etc., and inference is easy.

Unfortunately, the homogeneous random graph model fits basically no real networks. It implies a binomial (or, for large n, Poisson) degree distribution, and degree distributions are right-skewed; it implies very few triangles, which are abundant in real social networks; etc., etc. Since the model has two ingredients, independence of dyads and homogeneity of dyads, we might as well start by breaking *one* of them, and see if that helps. We'll start with homogeneity.

3 Inhomogeneous random graphs

An inhomogeneous random graph of size n is defined by the following two properties:

1. All dyads are mutually independent, $A_S \perp\!\!\!\perp A_T$ unless $S \cap T \neq \emptyset$;
2. There is a $n \times n$ matrix $\mathbf{p}$ where $\Pr\left(A_{ij} = 1\right) = p_{ij}$. We may call $\mathbf{p}$ the **expected adjacency matrix**, or **edge-probability matrix**.

A homogeneous random graph model is just one where all the entries in $\mathbf{p}$ are the same, say p. We have thus strictly increased the expressive power of the models. In doing so, it should be clear that we can fix some of the ways in which homogeneous random graphs fail to fit real networks. The expected degree of node i, for instance, is $\sum_j p_{ij}$, so we can ensure that there is heterogeneity in degree[3]. Triangles can be arranged by ensuring that for some trios of nodes i, j, k, p_{ij}, p_{jk} and p_{ik} are all above average. And so on and so forth.

The properties of inhomogeneous random graphs could be developed along much the same lines as homogeneous random graphs, but that is not the point of this course. Rather, what we need to notice is that we have, in a way, expanded the expressive power of the model *too much* — at least if we want to connect it to data. After all, every entry in $\mathbf{p}$ can be arbitrarily different from every other entry, without any violation of the defining properties. And this means that there is exactly *one* observation which is relevant to each p_{ij}, namely A_{ij}. With one observation for each parameter, we are obviously in a very bad position to do inference. We *could* do inference if instead of one graph with its adjacency matrix $\mathbf{A}$, we say may graphs, say m, with adjacency matrices $\mathbf{A}^{(1)}, \ldots \mathbf{A}^{(m)}$. If these were independent realizations of

[2] If we're dealing with simple graphs, which do not allow for self-edges, $A_{ii} = 0$ necessarily, and it would be more natural to make the denominator $n(n-1)$. Since $\frac{n^2}{n(n-1)} \to 1$, this makes no difference asymptotically.

[3] Actually, that just says there should be heterogeneity in *expected* degree. Call the (random) degree of node i K_i. Applying the bounded difference inequality (see Appendix A), we have that $\Pr\left(|K_i/n - \mathbb{E}\left[K_i/n\right]| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}$, so actual degrees will, with exponentially high probability, be close to expected degrees. (More exactly, fluctuations in degree which are bigger than $O(\sqrt{n})$ will be exponentially unlikely.)

the same inhomogeneous-random-graph process, we'd have

$$\frac{1}{m}\sum_{t=1}^{m} A_{ij}^{(t)} \rightarrow p_{ij} \tag{2}$$

by the law of large numbers, and so

$$\frac{1}{m}\sum_{t=1}^{m} \mathbf{A}^{(t)} \rightarrow \mathbf{p} \tag{3}$$

While there are some circumstances where it is reasonable to think of getting multiple independent graphs on the same set of nodes, it's not the usual situation in which we try to do inference.

4 Structured, inhomogeneous random graphs

If we are to be able to draw inferences about $\mathbf{p}$ from a *single* observed graph and its adjacency matrix $\mathbf{A}$, then $\mathbf{p}$ has to have some structure which means that multiple entries in $\mathbf{A}$ provide information on each entry in $\mathbf{p}$. Heuristically, $\mathbf{A}$ contains (at most) $O(n^2)$ degrees of freedom[4], so the number of free parameters determining $\mathbf{p}$ should really be $o(n)$ if inference is to have any hope of being consistent.

Here is a simple example to show that such inference-enabling structure can exist. Suppose that $p_{ij} = p$ if i and j are both even or both odd, but that $p_{ij} = q$ if i is even and j is odd or vice-versa. Then averaging A_{ij} over the even-even and odd-odd dyads yields a consistent estimate of p:

$$\frac{1}{n^2/2}\sum_{k=0}^{\lfloor n/2 \rfloor}\sum_{l=0}^{\lfloor n/2 \rfloor} A_{2k2l} + A_{2k+12l+1} \rightarrow p \tag{4}$$

Similarly, averaging edge indicators over the odd-even and even-odd pairs yields a consistent estimator for q. Clearly, this example is somewhat contrived, but it establishes that inference is possible; "now we're just haggling over the price".

4.1 Models based on nodal attributes

An important class of model of this sort is one based on node-level variables or attributes, say z_i for node i. Specifically, we assume the following:

1. All dyads are mutually independent, $A_S \perp\!\!\!\perp A_T$ unless $S \cap T \neq \emptyset$;
2. There is a $n \times n$ matrix $\mathbf{p}$ where $\Pr\left(A_{ij} = 1\right) = p_{ij}$.

[4]There are at most $n(n-1)$ non-trivial entries in the adjacency matrix for a directed graph, and only $n(n-1)/2 = \binom{n}{2}$ for an undirected. I say "at most" because if the sequence of graphs is known to be sparse, so that the number of edges is $O(n)$, then there are indeed only $O(n)$ degrees of freedom in the adjacency matrix.

3. There are variables z_i, taking values in $\mathcal{Z}$, and a function $w : \mathcal{Z} \times \mathcal{Z} \mapsto [0,1]$, such that $p_{ij} = w(z_i, z_j)$.

It is the **link-probability function** w which gives the structure to the inhomogeneous random graph.

Assuming we can observe the z_i, we can treat it just like any other feature, attribute or covariate. Since $\mathbb{E}\left[A_{ij}\right] = w(z_i, z_j)$, and the A_{ij} are mutually independent, inference on the function w is just like any other sort of inference for a regression function. Here are some examples.

Block model Suppose that all z_i take values in a finite, discrete set of categories, which without loss of generality we may take to be the integers from 1 to k. Introduce a $k \times k$ matrix $\mathbf{b}$, all of whose entries are in $[0,1]$. The classical **block model** then has

$$p_{ij} = b_{z_i z_j} \tag{5}$$

(In the older terminology of social network analysis [as in e.g. Wasserman and Faust 1994], this is "stochastic *a priori* block model".)

Inference is clearly possible here. To estimate b_{uv}, we need only take all (i,j) where $z_i = u$ and $z_j = v$ and average A_{ij}. Since such A_{ij} are, by hypothesis, IID binary variables, that average will converge on b_{uv} as the number of dyads in question tends to infinity.

Degree-corrected block model Suppose that $z_i = (u_i, r_i)$, where u is categorical as before, but r_i is a non-negative real number, and that

$$p_{ij} = b_{u_i u_j} r_i r_j \tag{6}$$

If i and j are in the same block, $u_i = u_j$, then the ratio of their expected degrees is the ratio of their rs, $\mathbb{E}\left[K_i\right] / \mathbb{E}\left[K_j\right] = r_i / r_j$. For this reason, this model is often called the **degree-corrected block model** (Karrer and Newman, 2011). This model does not really *explain* heterogeneity in degree within a block, it just takes it as a brute fact (encoded in the r_i) and proceeds from there.

This model is not, as it stands, identified. The problem is that if we take all the nodes where $u_i = u$ and multiply their r_i by the same factor, say c, we can compensate by dividing all the b_{uv} by c and nothing will change[5]. There are various possible ways of fixing this, such as requiring that r_i average to 1 for all the nodes in the same block. (See Karrer and Newman 2011 again.)

"Sociality" models If the z_i are just positive real numbers, we can have the model where

$$p_{ij} = z_i z_j \tag{7}$$

[5]Except for b_{uu}, which should be divided by c^2.

and so expected degree is proportional to z_i. This is, so to speak, the degree-corrected version of the undirected Erdos-Renyi model[6]. A different version of the same idea is to let z_i be an arbitrary real number, and set

$$p_{ij} = \frac{e^{z_i + z_j}}{1 + e^{z_i + z_j}} \tag{8}$$

or equivalently

$$\text{logit}\, p_{ij} = z_i + z_j \tag{9}$$

See exercise 2 for a comparison of these two parameterizations (which, working from memory, I got somewhat wrong in class).

"Social distance" models If $\mathcal{Z}$ is a multi-dimensional metric space, with each coordinate measuring some attribute of the node, we might naturally suppose that similar nodes are more likely to be connected, so that edge probabilities should depend on distance in this social space, sometimes called **Blau space**[7]. With a metric d on the $\mathcal{Z}$ space, a natural model would be

$$p_{ij} = w(z_i, z_j) = f(d(z_i, z_j)) \tag{10}$$

for some non-increasing function f of the distance. A popular model of this form comes from Hoff *et al.* (2002), where $\mathcal{Z}$ is taken to be $\mathbb{R}^p$ for some p, the metric is the ordinary Euclidean distance, and

$$w(z_i, z_j) = \overset{-1}{\text{logit}}\, \beta_0 - ||z - i - z_j|| \tag{11}$$

(See Exercise 3 for the reason why there is only an intercept and not a slope on the right hand side of this model.)

Obviously models of this sort could be made more complicated, e.g., by "degree correction" or "sociality" along the lines sketched above, or by weighting distance along some dimensions of Blau space more heavily than others, or even favoring heterophily along some directions, etc., etc.

4.2 Reduction of estimation to regression

If we can observe the z variables, then, at least in principle, the problem of estimating these models simply reduces to that of regression. This is because $w(z_i, z_j)$ is the expected value of A_{ij}, and regression just is the problem of inferring conditional expectation functions. If $\mathcal{Z} \times \mathcal{Z}$ is a metric space, for example, nearest-neighbor regression

[6]For a directed model, it would be natural to make z_i two-dimensional, $z_i = (z_i^{(1)}\ z_i^{(2)})$, and have $p_{ij} = z_i^{(1)} z_j^{(2)}$. We would also have to decide on whether the edge from i to j was dependent on the edge from j to i, i.e., on "reciprocity". This adds a certain amount of notational complexity without really changing the fundamental ideas, so I am skipping over directed graphs.

[7]The reference is to Blau (1977), though it's not clear to me that this is really what Blau had in mind when he talked about "[s]ocial structure [as] the distributions of a population among social positions in a multidimensional space of positions", or theorized about the conditions which would promote or retard homophily.

is universally consistent, i.e., it will eventually learn any not-too-pathological regression function (Györfi *et al.*, 2002, ch. 6). If we know, or are willing to guess, a specific parametric form, then doing parametric regression modeling will improve our efficiency. Of course estimation doesn't *have* to be handled by regression for models like this, but it *could* be.

5 Conditionally-independent dyad models

Unfortunately, we are rarely in the nice position of observing, *completely*, all the variables which are related to network tie formation. Variables which are not observable are **hidden** or **latent**. There are two common strategies for dealing with such variables:

1. Treat them as fixed but unknown parameters, and estimate them.
2. Treat them as random variables, and estimate their distribution.

When statisticians talk about "latent-variable models", they typically have the latter possibility in mind.

In the case of network models, if the node-level variables are treated as random, so that the random variable for node i is Z_i, then we have what are called **conditionally-independent dyad** (CID) models. The name arises from the fact that the models assume that $A_S \perp\!\!\!\perp A_T | Z$ for any non-intersecting sets of dyads S and T. These models further assume that $\Pr\left(A_{ij} = 1 | Z_i, Z_j\right) = w(Z_i, Z_j)$ for some function w, and that $Z_i \sim_{IID} \rho$ for some fixed distribution ρ.

Conditional on all the node-level latent variables, the probability of a graph is simple:

$$\Pr(\mathbf{A} = \mathbf{a} | Z = z) = \prod_{(i,j)} w(z_i, z_j)^{a_{ij}} (1 - w(z_i, z_j))^{1 - a_{ij}} \tag{12}$$

To get the unconditional probability of a graph, we need to multiply by the probability of seeing a particular configuration of the Zs, and sum over all possible configurations:

$$\Pr(\mathbf{A} = \mathbf{a}) = \sum_z \prod_{i=1}^{n} \rho(z_i) \prod_{(i,j)} w(z_i, z_j)^{a_{ij}} (1 - w(z_i, z_j))^{1 - a_{ij}} \tag{13}$$

You might well wonder whether, for instance, having IID draw from a fixed distribution for the Z_i is too restrictive; we will see in the next lecture that it is actually very natural.

All of the models given as examples of structured inhomogeneous random graphs become CID models once we make the node-level variables random. Block models become **stochastic block models**; social-space models become **continuous latent space models**.

5.1 Dependence of Dyads

Even though all dyads are conditionally independent given the Zs, they are unconditionally *dependent* (in general). The issue is that seeing whether or not one edge is present, say whether $A_{ik} = 1$ or not, gives us some information about the latent variables for the nodes involved (here, i and j), and that in turn gives us some information about their other edges (say, A_{ij}). If we take the model seriously as a generative story, then it's not that changing A_{ik} will have any *effect* on A_{ij}, but they are dependent because we can draw probabilistic inferences about A_{ij} from A_{ik}. (That is, we're dealing with Bayes's rule, rather than causality.)

Here is an easy example through which to see this. Suppose that Z_i is either $+1$ or -1 with equal probability[8], and that $w(z_i, z_j) = \text{logit}^{-1}(z_i + z_j)$. That is, there are two types of nodes, ones with lots of edges and ones with fewer edges. Then A_{ij} is dependent of A_{ik}. After all, if there is an edge between i and k, that has to make it more probable that Z_i is one of the lots-of-edges nodes, i.e., $\Pr(Z_i = +1 | A_{ik} = 1) > \Pr(Z_i = +1)$. But this in turn must increase the probability that there is an edge between i and j (even if $Z_j = -1$).

One might wonder whether A_{kl} and A_{ij} are dependent (i, j) and (k, l) share no nodes in common. An easy way to see that the answer is "no", at least when the Z_i are generated independently, is to draw the graphical model for the dependence among the random variables[9], as in Figure 1. Any path from A_{ij} to A_{kl} has to go through two Z variables, one from each dyad, say Z_j and Z_k, as in the figure. But to connect those variables, the path has to go through the edge indicator for a dyad they have in common, say A_{jk}. That edge indicator forms a collider on the path, and colliders block dependence *unless* they are conditioned on. Thus, unconditionally, $A_{ij} \perp\!\!\!\perp A_{kl}$. Since conditioning on colliders "activates" them, opening the path at that step, we also have $A_{ij} \not\perp\!\!\!\perp A_{kl} | A_{jk}$.

5.2 Inference and likelihood

One consequence of making the node variables Z_i latent is that the parameters to be inferred are those of their distribution, ρ, and the link probability function, w. If those follow parametric models, then we ought to have a fixed number of parameters, while still having $O(n^2)$ degrees of freedom in the data, and so inference should be possible. Even if we let some aspects of the model become non-parametric, it should still be possible to estimate them, provided we impose the sort of capacity control we're used to from other contexts, like regression or density estimation.

For likelihood-based inference, what we'd want to maximize is just Eq. 13. What makes things like this hard to maximize is the presence of the sum over possible values of Z. (Remember that the log of a sum is not a sum of logs.) This is generally intractable, but it's a common problem with latent-variable models, and as such

[8] Such random variables are sometimes called **Rademacher** random variables.

[9] For a review of the terminology of graphical models used here, see Koller and Friedman (2009), or Shalizi (forthcoming).

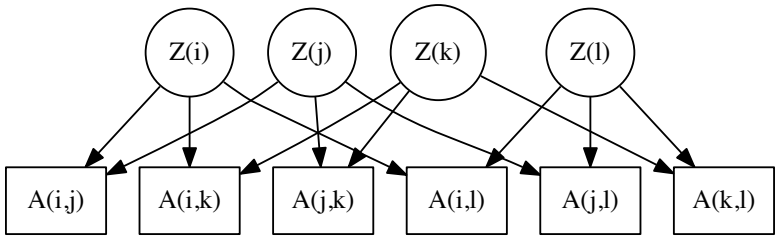

Figure 1: Graphical model depicting dependencies among node-level latent variables (Zs) and observable edge indicators (*A*s).

has a common solution, the EM algorithm. In the E ("expectation") step, we find $\Pr\left(Z|\mathbf{A}=\mathbf{a},\hat{\theta}\right)$ for some guess $\hat{\theta}$ about the parameters in ρ and w. In the M step ("maximization"), we find the likelihood-maximizing value of the parameters θ, for some guess about the distribution of Z conditional on $\mathbf{A}$. Alternating between these two steps keeps increasing the likelihood, until we get to a local maximum[10].

Computationally, the M step tends to be easy — if we knew Z, we'd just do regression. The E step tends to be where computation gets tricky, because the distribution of each Z_i has to be conditioned on the state of every dyad i participates in (i.e., i's edge or non-edge to every other j), *and* on the latent variables for all those other nodes. One successful strategy for dealing with this in stochastic block models is belief propagation (Decelle *et al.*, 2011b,a); another is to switch to some sort of pseudo-likelihood method (Amini *et al.*, 2013).

6 Permutation-invariance (exchangeability)

CID models have a very important symmetry property, which is that they lead to distributions over graphs which are invariant under permutation. To make this clear, we need a little notation.

By a **permutation**, we mean a 1-1, invertible mapping from the integers $1:n$ into themselves; a generic permutation will be π, and the number it takes $i \in 1:n$ to will be written $\pi(i)$. The permutations form a group, generated by the permutations which just exchange the position of a single pair of numbers. Given a random vector Z, the permuted vector Z^{π} is defined by $Z_i = Z^{\pi}_{\pi(i)}$; this implies $Z^{\pi}_i = Z_{\pi^{-1}(i)}$. Similarly, when we permute the nodes of a graph, the new adjacency matrix $\mathbf{A}^{\pi}$ will be given by $A_{ij} = A^{\pi}_{\pi(i)\pi(j)}$, or $A^{\pi}_{ij} = A_{\pi^{-1}(i)\pi^{-1}(j)}$.

[10]For a very lucid view of the EM algorithm, which in particular explains how it can be seen as maximizing a lower bound on the likelihood, see Neal and Hinton (1998).

A distribution μ over graphs is **permutation-invariant** when for every adjacency matrix $\mathbf{a}$, and every permutation π,

$$\mu(\mathbf{a}) = \mu(\mathbf{a}^{\pi}) \tag{14}$$

Such distributions are also called (finitely) **exchangeable**. Another way to think about it is this: two graphs are **isomorphic** if and only if there is some permutation of the nodes which preserves edges, i.e., $\mathbf{a}$ and $\mathbf{b}$ are isomorphic iff $\mathbf{b} = \mathbf{a}^{\pi}$ for some permutation π. Exchangeable or permutation-invariant distributions are ones which give equal probability to isomorphic graphs.

To see that every CID distribution is permutation-invariant, notice that under any ρ,

$$\rho(Z = z) = \prod_{i=1}^{n} \rho(Z_i = z_i) = \prod_{i=1}^{n} \rho(Z^{\pi} = z_i^{\pi}) = \rho(Z^{\pi} = z^{\pi}) \tag{15}$$

because the coordinates of Z are IID. Now similarly,

$$w(z_i, z_j)^{a_{ij}} = w(z^{\pi}_{\pi(i)}, z^{\pi}_{\pi(j)})^{a^{\pi}_{\pi(i)\pi(j)}} \tag{16}$$

and likewise

$$(1 - w(z_i, z_j))^{1-a_{ij}} = (1 - w(z^{\pi}_{\pi(i)}, z^{\pi}_{\pi(j)}))^{1-a^{\pi}_{\pi(i)\pi(j)}} \tag{17}$$

Substituting into Eq. 13 thus makes it clear that the whole distribution is permutation-invariant.

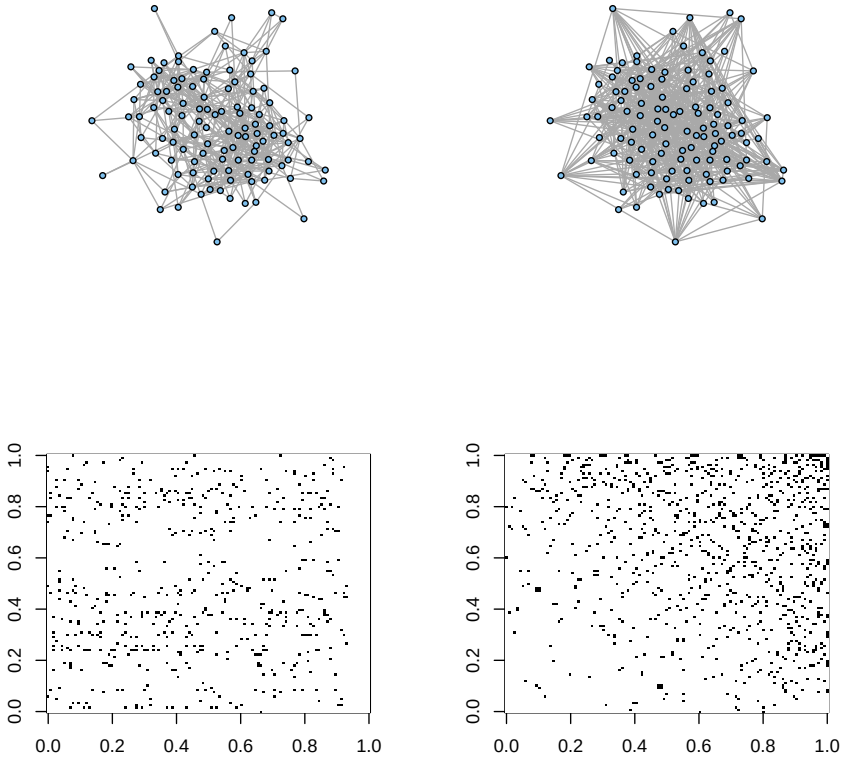

Figure 2: Graph layouts (above) and adjacency matrices (below) for the same (real) network in two permutations (left and right); the ultimate data source is Coleman *et al.* (1957). The left-hand column uses the nodes in the order given by the original data source, the right-hand column a "canonical" permutation (see `canonical.permutation` in R's `igraph` package (Csardi and Nepusz, 2006). In the adjacency-matrix plots, edges are indicated by black squares, non-edges by white. In a permutation-invariant (or "exchangeable") distribution over graphs, the graph on the left must have the same probability as the graph on the right, and as every other isomorphic graph.

7 Exercises

1. A sequence of graphs is **dense** if the number of edges is $O(n^2)$; it is **sparse** if the number of edges is $o(n^2)$. Suppose that the number of edges is $O(n)$, what is sometimes called "very sparse", and more specifically that the number of edges is cn for some $c > 0$.

   (a) Show that the number of graphs with n nodes and cn edges is

   $$\binom{\binom{n}{2}}{cn} \tag{18}$$

   (b) Find an asymptotic formula for the log number of such sparse graphs, in terms of n, c, and numerical constants. *Hint:* Stirling's formula.

   (c) Evaluate the following claim: "Describing a graph with n nodes and cn edges takes at most $O(n)$ bits".

2. Compare the model where $p_{ij} = z_i z_j$ to the one where $p_{ij} = \frac{e^{\beta_i+\beta_j}}{1+e^{\beta_i+\beta_j}}$.

   (a) Show that both these models include homogeneous random graphs as special cases. Can all the p_{ij} be equal if some of the z_i are not equal? Can all the p_{ij} be equal if some of the β_i are unequal?

   (b) Show that, *in general*, these two models are not equivalent, because there is no 1-1, invertible mapping from zs to βs which preserves all the edge probabilities. *Hint:* Suppose that we've found vectors z and β which produce the same $\mathbf{p}$ (which are not all zero). Check that replacing z_i by $z_i/2$ reduces p_{ij} by a factor of 2 in the first model, for all j, but leaves p_{kj} unchanged (unless k or $j = i$, of course). Show that no modification of β will reduce all the p_{ij} by a factor of 2, without changing p_{kj}.

   (c) Show that the two models coincide for sparse graphs, in the sense that as $z \to \vec{0}$, using $\beta_i = \log z_i$ in the second model provides an increasingly good approximation to the distribution of the first model, and vice-versa that as $\beta \to -\infty\vec{1}$, using $z_i = e^{\beta_i}$ gives an increasingly good approximation to the second model.

3. Eq. 11, for the model of Hoff *et al.* (2002), looks like a logistic regression, but with only an intercept coefficient, the slope being fixed at one. It might seem more natural to allow the slope to vary as well, so that

   $$\text{logit}\, w(z_i, z_j) = \beta_0 - \beta_1 ||z_i - z_j|| \tag{19}$$

   Prove that every model with $\beta_1 \neq 1$ is equivalent to another model where $\beta_1 = 1$, but z_i is replaced by $\beta_1 z_i$.

A Inequalities

This will probably be the first in a series of appendices on useful probabilistic inequalities.

A.1 Bounded difference inequality

Let $X_1, X_2, \ldots X_n$ be independent random variables, not necessarily identically distributed; indeed, they make take values in different spaces Ξ_i. Let f be any real-valued function of these variables, which has **bounded differences**: for each i, there is a constant c_i such that

$$\sup_{x_1,\ldots x_i,\ldots x_n,x_i'} |f(x_1,\ldots x_{i-1},x_i,x_{i+1},\ldots x_n) - f(x_1,x_{i-1},\ldots x_i',x_{i+1},\ldots x_n)| \leq c_i \quad (20)$$

for any $x_1 \in \Xi_1, \ldots x_n \in \Xi_n$, and $x_i, x_i' \in \Xi_i$. In words, changing the i^{th} input to f (and nothing else) can change the output of f by at most c_i. Then

$$\Pr\left(f(X_1,\ldots X_n) - \mathbb{E}[f] \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}} \quad (21)$$

This implies (why?) that

$$\Pr\left(|f(X_1,\ldots X_n) - \mathbb{E}[f]| \geq \epsilon\right) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}} \quad (22)$$

REFERENCE: Boucheron *et al.* (2013, Theorem 6.2, p. 171).

References

Amini, Arash A., Aiyou Chen, Peter J. Bickel and Elizaveta Levina (2013). "Pseudo-likelihood methods for community detection in large sparse networks." *Annals of Statistics*, **41**: 2097–2122. URL `http://arxiv.org/abs/1207.2340`.

Blau, Peter M. (1977). "A Macrosociological Theory of Social Structure." *American Journal of Sociology*, **83**: 26–54. URL `http://www.jstor.org/stable/2777762`. doi:10.1086/226505.

Boucheron, Stéphane, Gábor Lugosi and Pascal Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford University Press.

Coleman, James, Elihu Katz and Herbert Menzel (1957). "The Diffusion of an Innovation Among Physicians." *Sociometry*, **20**: 253–270. URL `http://www.jstor.org/stable/2785979`. doi:10.2307/2785979.

Csardi, Gabor and Tamas Nepusz (2006). *The igraph software package for complex network research*. URL `http://www.interjournal.org/manuscript_abstract.php?361100992`.

Decelle, Aurelien, Florent Krzakala, Cristopher Moore and Lenka Zdeborová (2011a). "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications." *Physical Review E*, **84**: 066106. URL http://arxiv.org/abs/1109.3041. doi:10.1103/PhysRevE.84.066106.

— (2011b). "Inference and Phase Transitions in the Detection of Modules in Sparse Networks." *Physical Review Letters*, **107**: 065701. URL http://arxiv.org/abs/1102.1182. doi:10.1103/PhysRevLett.107.065701.

Erdös, P. and A. Rényi (1960). "On the Evolution of Random Graphs." *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, **5**: 17–61. Reprinted (Newman *et al.*, 2006, pp. 38–61).

Gilbert, E. N. (1959). "Random Graphs." *Annals of Mathematical Statistics*, **30**: 1141–1144. doi:10.1214/aoms/1177706098.

Györfi, László, Michael Kohler, Adam Krzyżak and Harro Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer-Verlag.

Hoff, Peter D., Adrian E. Raftery and Mark S. Handcock (2002). "Latent Space Approaches to Social Network Analysis." *Journal of the American Statistical Association*, **97**: 1090–1098. URL http://www.stat.washington.edu/research/reports/2001/tr399.pdf.

Karrer, Brian and Mark E. J. Newman (2011). "Stochastic Blockmodels and Community Structure in Networks." *Physical Review E*, **83**: 016107. URL http://arxiv.org/abs/1008.3926.

Koller, Daphne and Nir Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, Massachusetts: MIT Press.

Neal, Radford M. and Geoffrey E. Hinton (1998). "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants." In *Learning in Graphical Models* (Michael I. Jordan, ed.), pp. 355–368. Dordrecht: Kluwer Academic. URL http://www.cs.toronto.edu/~radford/em.abstract.html.

Newman, Mark, Albert-László Barabási and Duncan J. Watts (eds.) (2006). *The Structure and Dynamics of Networks*, Princeton, New Jersey. Princeton University Press.

Shalizi, Cosma Rohilla (forthcoming). *Advanced Data Analysis from an Elemntary Point of View*. Cambridge, England: Cambridge University Press. URL http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV.

Solomonoff, Ray and Anatol Rapoport (1951). "Connectivity of Random Nets." *Bulletin of Mathematical Biophysics*, **13**: 107–117. Reprinted (Newman *et al.*, 2006, pp. 27–37).

Wasserman, Stanley and Katherine Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge, England: Cambridge University Press.