Lecture 2: Exchangeable networks and the Aldous-Hoover representation theorem

36-781: Advanced Statistical Network Models
Mini-semester II, Fall 2016
Instructor: Cosma Shalizi

Scribe: Momin M. Malik

27 October 2016

Contents

1 CID models (again)

Recall: conditionally independent dyad (CID) models. Each node i, $i \in \{1, ..., n\}$, has its own random variable $Z_i \in \mathcal{Z}$ ($\mathcal{Z}$ could be integers, reals, qualitative set of labels, hyperbolic space, anything we like), drawn iid from the same distribution, call it ρ. Let $Z = (Z_1, ..., Z_n)$ be the vector of the Z_i's.

Assert: $Pr(A_{ij} = 1|Z_1, ..., Z_n) = w(Z_i, Z_j)$ for some w: the tie probability is a function of the two node probabilities alone. That is, all dyads are independent when conditioned on relevant node variables,

$$A_{ij} \perp\!\!\!\perp A_{kl}|Z_i, Z_j, Z_k, Z_l.$$

In general, edges will not be marginally independent, but they will be conditionally independent.

Hierarchically (for plugging into WINBUGS or STAN):

$$Z_i \overset{\text{iid}}{\sim} \rho$$
$$A_{ij} \overset{\text{iid}}{\sim} \text{Bern}(w(Z_i, Z_j))$$

Start by generating the Z_i's independently, then each edge indicator variable conditioned on the Z_is has a Bernoulli distribution, with the probability being a function of Z_i and Z_j.

2 Permutation and permutation invariance ("exchangeability")

Properties of this model. We introduce idea of permuting the node labels. If we have n nodes, a permutation is $\pi : \{1, ..., n\} \mapsto \{1, ..., n\}$ with an inverse π^{-1}. Notation: for a vector Z, Z^π will be the permuted vector, so $Z_i^\pi = Z_{\pi^{-1}(i)}$. The ith entry in Z^π is equal to the corresponding entry in Z for whatever position in Z got mapped to i. The other way to read that is, $Z_i = Z^\pi_{\pi(i)}$.

Similarly, for a matrix $\mathbf{a}$, permute it to $\mathbf{a}^\pi$, which will be the matrix with both rows and columns permuted by π. This means that

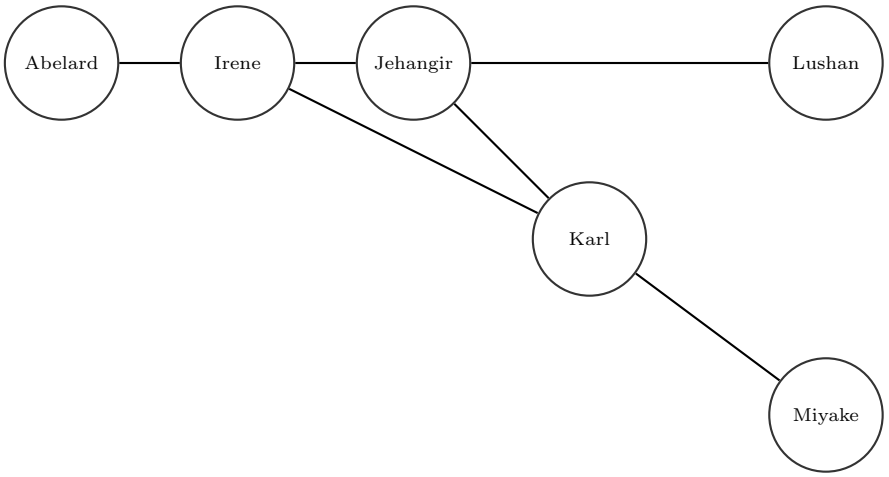
$$a_{ij} = a^\pi_{\pi(i),\pi(j)}$$

or, equivalently,

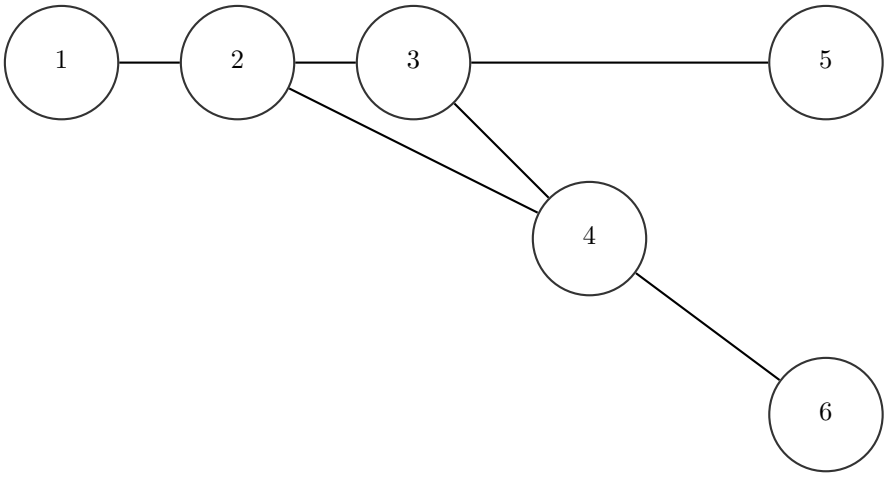
$$a^\pi_{ij} = a_{\pi^{-1}(i),\pi^{-1}(j)}.$$

Definition. *Permutation-invariance* of the distribution holds when $Pr(\mathbf{A} = \mathbf{a}) = Pr(\mathbf{A} = \mathbf{a}^\pi)$ for any permutation π and any adjacency matrix $\mathbf{a}$. That is, permuting the adjacency matrix doesn't change its probability. Then we say that the distribution is permutation-invariant.

What does it mean, and why do we care?



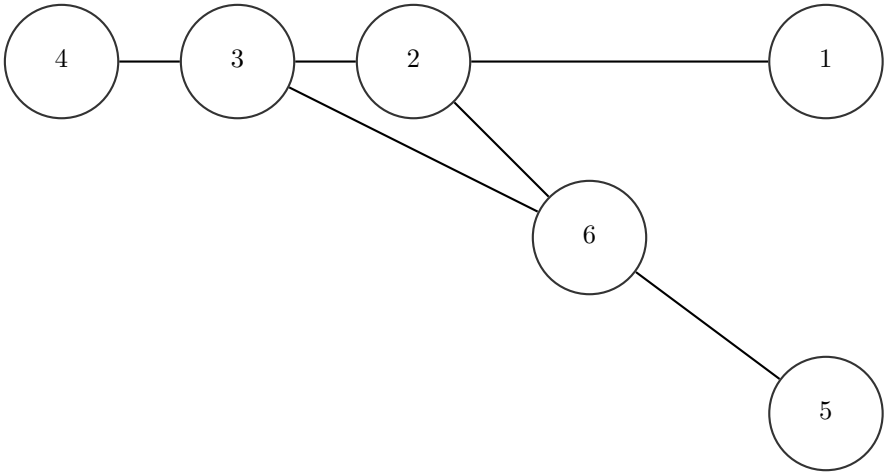
Since Cosma was raised with the Latin alphabet, it is natural for him to order them left to right, top to bottom:



This corresponds to the adjacency matrix:

.	1	.	.	.	.
1	.	1	1	.	.
.	1	.	1	1	.
.	1	1	.	.	1
.	.	1	.	.	.
.	.	.	1	.	.

Given the way it was drawn and the way Cosma writes, this is a perfectly natural ordering. But Cosma's grandfather was brought up speaking Persian, so his natural ordering might be something like this:



The adjacency matrix of this would look quite different:

.	1	.	.	.	.
1	.	1	.	.	1
.	1	.	1	.	1
.	.	1	.	.	.
.	.	.	.	.	1
.	1	1	.	1	.

But the only thing that has changed is the order of the nodes. The order in which we write things down is usually completely arbitrary, and irrelevant to whatever is going on in the world that makes these graphs. Our probability model should respect this.

We should be able to *exchange* nodes without changing probabilities. Exchangeable = permutation-invariant. Seems like a very natural and harmless thing to require. To the end of the course, we will take about some reasons why you may not want exchangeability since it leads to weird things, and what you might want to do instead. This seems natural and harmless but has some non-trivial consequences which is what got people excited about these models.

3 CID $\implies$ exchangeable

Claim: Every CID model is exchangeable. $\forall$ adjacency matrices $\mathbf{a}$ and $\forall$ permutations π,

$$Pr(\mathbf{A} = \mathbf{a}) = Pr(\mathbf{A} = \mathbf{a}^\pi)$$

Proof: Since Z_i's are IID, $Z_i \overset{\text{iid}}{\sim} \rho$, $p(z_1, ..., z_n) = \prod_{i=1}^n p(z_i)$. And, since $A_{ij} \overset{\text{iid}}{\sim} \text{Bern}(w(z_i, z_j))$,

$$Pr(\mathbf{A} = \mathbf{a}) = \sum_{z_1,...,z_n} \prod_{i=1}^n p(z_i) \prod_{(i,j)} w(z_i, z_j)^{a_{ij}} (1 - w(z_i, z_j))^{1-a_{ij}} \tag{1}$$

First notice, we can swap around z_i's and it won't change the probability, $Pr(Z_{1:n} = z_{1:n}) = \prod_{i=1}^n p(z_{\pi(i)})$. Then,

$$Pr(Z_{1:n} = z_{1:n}) = Pr(Z_{1:n} = z^\pi_{1:n})$$

So the first part, $\prod_{i=1}^n p(z_i)$, will be invariant under permutation.

Now, if we look at $w(z_i, z_j)$, it will change under permutation. But what we care about is $w(z_i, z_j)^{a_{ij}}$.

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

We are plugging the same numbers into w, so we get the same number out, and we raise the same numbers to the same power, so they must be equal. Apply the same argument to the $(1 - w(z_i, z_j))^{1-a_{ij}}$ factor.

Every term in the product in eqn. (1) is left alone by permutation. And since the whole product is left alone by permutation, the whole sum is left alone by permutation. Thus,

$$\begin{aligned} Pr(\mathbf{A} = \mathbf{a}) &= \sum_{z_1,...,z_n} \prod_{i=1}^n p(z_i) \prod_{(i,j)} w(z_i, z_j)^{a_{ij}} (1 - a(z_i, z_j)^{1-a_{ij}} \\ &= Pr(\mathbf{A} = \mathbf{a}^\pi) \ \square \end{aligned}$$

Claim 1: Every CID model is exchangeable,

$$Pr(\mathbf{A} = \mathbf{a}) = Pr(\mathbf{A} = \mathbf{a}^\pi).$$

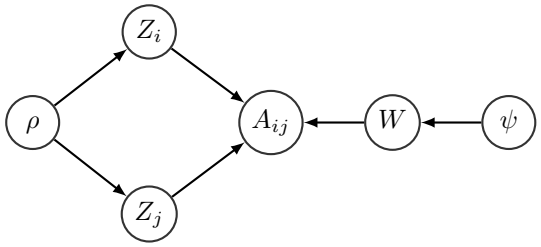
4 Mixture of CID $\implies$ exchangeable

Claim 2: If a model is a mixture of CID models, it's still exchangeable.

A mixture of CID models just adds one more level to the hierarchical structure. Before, we had some fixed function w. Now say, we draw that function from some distribution, $W \sim \psi$. In 36-752, Advanced Probability Theory, you learn how to do probability over infinite spaces including spaces of functions, but just trust that it works for now.

$$\begin{aligned} W &\sim \psi \\ Z_i|W &\overset{\text{iid}}{\sim} \rho \\ A_{ij}|Z, W &\overset{\text{iid}}{\sim} \text{Bern}(w(Z_i, Z_j)) \end{aligned}$$

As a graphical model,



Example of mixture of CID models: Take some $k \times k$ matrix of probabilities, $\mathbf{B}$. Assign nodes to blocks, and conditional on node assignment and the random affinity matrix, get the probability of an edge between two nodes.

1. Choose a random block affinity matrix $\mathbf{B}$.
2. Assign nodes to blocks with distribution ρ, so Z_i is the block assignment of node i.
3. $Pr(A_{ij} = 1|Z_1, ..., Z_n, \mathbf{B}) = B_{Z_i, Z_j} = W(Z_i, Z_j)$.

Proof of Claim 2: We want to show $Pr(\mathbf{A} = \mathbf{a}) = Pr(\mathbf{A} = \mathbf{a}^\pi)$. This is a marginal probability: we have latent random variables, Z's and W's lurking around. So I can certainly rewrite this probability,

$$Pr(\mathbf{A} = \mathbf{a}) = \sum_w Pr(\mathbf{A} = \mathbf{a}, W = w)$$

For the permuted one as well,

$$Pr(\mathbf{A} = \mathbf{a}^\pi) = \sum_w Pr(\mathbf{A} = \mathbf{a}^\pi, W = w)$$

Then we are asking,

$$\begin{aligned} Pr(\mathbf{A} = \mathbf{a}) &\overset{?}{=} Pr(\mathbf{A} = \mathbf{a}^\pi) \\ \sum_w Pr(\mathbf{A} = \mathbf{a}, W = w) &\overset{?}{=} \sum_w Pr(\mathbf{A} = \mathbf{a}^\pi, W = w) \\ \sum_w Pr(\mathbf{A} = \mathbf{a}|W = w)Pr(W = w) &\overset{?}{=} \sum_w Pr(\mathbf{A} = \mathbf{a}^\pi|W = w)Pr(W = w) \end{aligned}$$

For any w, by Claim 1, $Pr(\mathbf{A} = \mathbf{a}|W = w) = Pr(\mathbf{A} = \mathbf{a}^\pi|W = w)$. Once we condition on random W, we get back the same thing.

Proof of Claim 2, reprise:

$$\begin{aligned} Pr(\mathbf{A} = \mathbf{a}) &= \sum_w Pr(\mathbf{A} = \mathbf{a}, W = w) \\ &= \sum_w Pr(\mathbf{A} = \mathbf{a}|W = w)Pr(W = w) \\ &= \sum_w Pr(\mathbf{A} = \mathbf{a}^\pi|W = w)Pr(W = w) \qquad \text{(Claim 1)} \\ &= \sum_w Pr(\mathbf{A} = \mathbf{a}^\pi, W = w) \\ &= Pr(\mathbf{A} = \mathbf{a}^\pi) \end{aligned}$$

You might imagine that we should have $\rho(w)$ in this derivation; it might make interpretation easier, or make the model more natural, but we will see that this is not mathematically necessary. We can always get away with using a single ρ distribution no matter the w, rather than making it change with the w.

5 Exchangeable $\implies$ mixture of CID

Claim 3: If μ is a exchangeable distribution over graphs for all n, then μ is a mixture of CID models (or a CID model).

Unfortunately, does not have such a nice proof. All the proofs are extremely long, and require extremely advanced probability, so we won't go through. This is a special case of a result called the Aldous-Hoover Theorem.[1]

Theorem (Aldous-Hoover): Fix any set of adjacency matrices S (any set of graphs). $\mu(\mathbf{A} \in S)$ is the probability, under the distribution μ, that $\mathbf{A}$ is in the set S. Then there exists a distribution ψ over W's, and a fixed distribution of Z_i's, ρ (independent of W), such that

$$\mu = \sum_w \psi(w)Pr(\mathbf{A} \in S|W = w)$$

(Note: $Pr(\mathbf{A} \in S|W = w)$ is the CID model). We can always write the mixture as the probability of some CID models, each weighted by the probability of picking that w in the first place. You can always use the same distribution of the Z_i's in the CID models.

Claim 3: If a model is exchangeable for all n, then it's either a CID model or a mixture of CID models.

It's annoying that the proof in one direction is simple, and in the other direction it got Kallenberg tenure, but such is the nature of mathematics.

[1] David Aldous proved a version of this theorem for "row-column exchangeable" or "separately exchangeable" arrays, where (naturally) we get to apply separate permutations to the rows and the columns, in Aldous (1981), a fairly comprehensible and quite "probabilistic" paper. What we need for networks is a result where the *same* permutation must be applied to the rows and to the columns of the adjacency matrix, i.e., a result for "jointly exchangeable" arrays. Here priority goes to an unpublished 1979 technical report at the Institute for Advanced Study by Douglas Hoover, which Cosma has never seen, but is reported to use a deep branch of mathematical logic called model theory. There is a much more comprehensible proof by Olav Kallenberg, given in chapter 7 of his book which is in the optional readings (Kallenberg, 2005). It uses some more advanced probability-theory notions, but nothing that fundamentally goes beyond 752. Kallenberg's proof is about 20 pages, the first 15 of which is just set up. We could spend the course just going through the proof, so Cosma is asking us to take on faith that we could work through those 20 pages and see that Kallenberg got everything right.

At this point, we invoke another probability fact without proof, but hopefully it is more plausible intuitively:

Claim 4: If Z is a random variable on any reasonable sample space $\mathcal{X}$, then there exists some function $f : [0,1] \mapsto \mathcal{X}$ such that $X = f(u)$ where $U \sim$ Uniform(0,1). If I want to generate a random variable on some arbitrary sample space (so long as it's not some horror beloved by mathematicians), all I need to do is generate a uniformly distributed number between 0 and 1, there is some function that will transform that to X. ("Reasonable" includes Borel spaces, so: finite sets, $\mathbb{R}$, $\mathbb{R}^d$, etc., any sample space you would naturally choose to work in).

Then f is not unique. If $\phi : [0,1] \mapsto [0,1]$ with ϕ^{-1} existing and ϕ preserves length, then $\phi(U) \sim$ Uniform(0,1). That is, we can map the unit interval to itself. There are infinitely many ways to do this that preserve length and have inverses. So I could equally well have written $X = f(\phi(u)) = g(u)$. Again, we won't prove: being precise about the limits of "reasonable" here is what led to people developing a lot of measure theory in the early 20th century.

Claim 5: Any CID model can be written in terms of n uniform distributions for nodes and 1 uniform distribution per dyad. That is, we don't have to worry about weird space $\mathcal{Z}$ in which each node lived before.

$$\begin{aligned} Z_i &= f(U_i) \quad \text{where } U_i \overset{\text{iid}}{\sim} \text{Unif}(0,1) \\ A_{ij} &= h_{ij}(U_{ij}) \quad \text{where } U_{ij} \overset{\text{iid}}{\sim} \text{Unif}(0,1) \end{aligned}$$

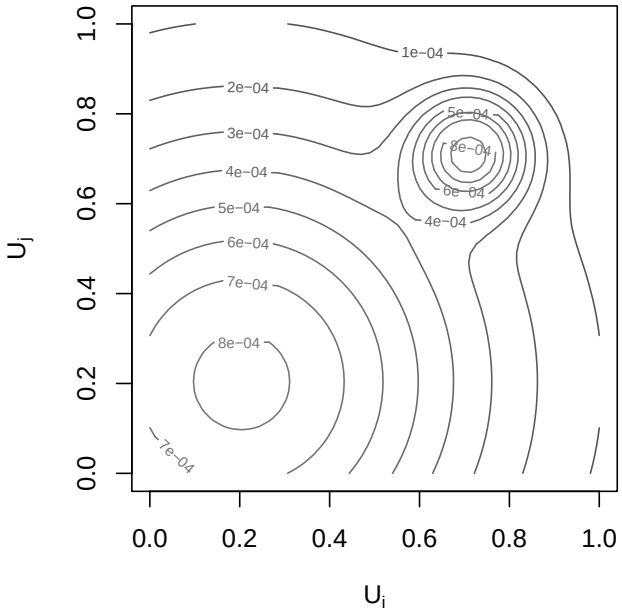
A_{ij} is binary and 1 with a certain probability, so this function can be taken to be the indicator for whether $U_{ij} \leq w(Z_i, Z_j)$, that is,

$$U_{ij} \overset{\text{iid}}{\sim} \text{Unif}(0,1) = \mathbb{1}\{U_{ij} \leq w(Z_i, Z_j)\}$$

This is not the only possible representation, but it's simply to compare a uniform deviate to a threshold to determine whether an edge is present.

When I have $w(Z_i, Z_j)$, there is some f that makes this work as a function of random noise, $w(Z_i, Z_j) = w(f(U_i), f(U_j)) = w'(U_i, U_j)$. Then, $A_{ij} = \mathbb{1}\{U_{ij} \leq w'(U_i, U_j)\}$. So in any CID model, we can always take the node variables instead of the Z's to be independent uniforms on $[0,1]$, i.e., we can always take ρ for the locations to be Unif$(0,1)$.

Therefore, we can always take w to be a function $w : [0,1] \times [0,1] \mapsto [0,1]$, a function from the unit square to the unit interval. This mapping might look something like this:



6 "Functional" representations

6.1 Functional representation of CID models

We can write any CID model as a function of uniformly distributed noise. For historical reasons, this is called a *functional representation*.

$$\begin{aligned} U_1, ..., U_n &\overset{\text{iid}}{\sim} \text{Unif}(0,1) \\ U_{ij} &\overset{\text{iid}}{\sim} \text{Unif}(0,1) \\ A_{ij} &= \mathbb{1}\{U_{ij} \leq w(U_i, U_j)\} \end{aligned}$$

6.2 Functional representation of mixture

Add one more noise variable, uniform U_0 which picks $w = f(U_0)$. Then,

$$A_{ij} = \mathbb{1}\{U_{ij} \leq f(U_0, U_i, U_j)\}$$

and we can represent any mixture of CID as transformations only of random noise.

So long as we believe the graph is exchangeable, this is what the model has to look like, or something similar to this. Much of what we will explore later in the course is how and under what conditions we can estimate models of this form, and what we can do once we are willing to believe that the model has to look like this.

7 Can we learn the representation?

There is a step before that, which is required, which is that we haven't really said, working at this level of generality: can we figure out w? If it has a specific parametric form, we can use EM; but, can we learn an arbitrary w? To do that, we have to look at what happens to graphs as they get very large: we have to look at limits of graphs. The limit gives us lots of information about what w has to look like. For that, on Tuesday, we need to talk about how we can have limits of graph sequences; on Thursday, we will see how they line up with w's.

References

Aldous, David J. (1981). "Representations for partially exchangeable arrays of random variables." *Journal of Multivariate Analysis*, **11**: 581–598. doi:10.1016/0047-259X(81)90099-3.

Kallenberg, Olav (2005). *Probabilistic Symmetries and Invariance Principles*. New York: Springer-Verlag.