1 More models

1.1 The Preferential Attachment Model (PAM)

This very popular model was proposed by Albert and Barabasi. It is a generative model for both nodes and edges, so it can be thought of as a graph process.

Algorithm PreferentialAttachment

Input number of nodes n, new node degree d

Initialize with $V = \{1\}$, $E = \emptyset$.

for $i = 2, \dots n$

1. create new node i
2. select up to d nodes $\{j_1, \dots j_d\} \subset V$ at random, with probability proportional to their degrees, without replacement. (i.e. renormalize the probability after each node is selected).
3. add edges ij_l, $l = 1 : \min(d, |V|)$ to E and add i to V

This model explains well the heavy tailed-distribution of node degrees observed in real networks. Variations have been proposed to account for other characteristic of real networks, like the slow-growing (or bounded ?) diameter.

1.2 "Expected degree" models

This class of models removes the assumptions that nodes are all alike. Herenodes are allowed to have different **weights** (or **propensities**) modulating their probability to form links with other nodes.

Model and parameters The simplest version of this model is an "augmented SBM" defined as follows: V is the set of nodes, K the number of disjoint communities, $B \in [0,1]^{K \times K}$ a matrix of link strenghts between communities (symmetric). Each node i has a paramer p_i, measuring its propensity to form ties, and $z_i \in \{1, \dots K\}$ is the cluster containing node i. The probability of edge ij is given by

$$P_{ij} = p_i p_j B_{z_i z_j} \tag{1}$$

One must choose parameters so that the above does not exceed 1.

A variant with overlapping communities (from Arora and Ge) Set of nodes V is given, together with a set of communities $C_{1:K} \subset V$, which may have non-empty intersections. Each node i may belong to up to d communities. For each community C, the **propensity of i in C** is $p_{iC} \in (0, 1]$. For every pair of nodes $i \neq j$ in V, the edge probability P_{ij} is

$$P_{ij} \geq \max_{C: i,j \in C} p_{iC} p_{jC} \text{ if } i, j \text{ have some } C \text{ in common} \tag{2}$$

Besides the intra-community edges, one assumes that there are edges into C from nodes outside C. These edges are treated as outliers; one assumes that each node $i' \notin C$ has no more than $(\alpha - \varepsilon)|C|$ edges into C, with $\alpha = \min_{i \in C} p_{iC}^2$. Hence, there is a gap between $i' \notin C$ and $i \in C$ w.r.t the number of edges into C. Another assumption is that a fraction $\gamma > 0$ of a node's edges are intra-communitty.

With these conditions, an algorithm exists that recovers all the communities in time $\mathcal{O}(n^{C_{\alpha,\gamma,\varepsilon} \max |C| d})$ (this is called quasi-polynomial time, i.e. polynomial in n if $\max C$ can be considered bounded).

2 A spectral algorithm for SBM estimation

This algorithm is simple and elegant. It is essentially a spectral clustering algorithm. It does not assume that the number of communities is known.

Algorithm Spectral community detection (Rohe,Chatterjee,Yu)
Input Graph adjacency matrix Y and node set V, $|V| = n$

1. Denote by d_i the degree of node $i \in V$ and by $D = \text{diag}\{d_i,\, i \in V\}$.
2. Compute the Laplacian $L = D^{-1/2} Y D^{-1/2}$.
3. Compute the non-zero eigenvalues of L, ordered by magnitude $|\lambda_1| = 1 \geq |\lambda_2| \geq \ldots \geq |\lambda_K| > 0$. The number of clusters K is the number of non-zero eigenvalues.
   Compute also the eigenvector matrix $X \in \mathbb{R}^{n \times K}$ associated with theses eigenvalues.
4. Now use K-means to cluster the n rows of X as points in $\mathbb{R}^K$.
   Return the resulting clustering.

Remarks:

- The actual normalized Laplacian is not L, but $I - L$, which has the same eigenvectors as L.

- Once the cluster assignments are obtained, the estimation of B is straightforward (assuming the assignments are correct).
- Exercise What are the differences between this algorithm and "standard" spectral clustering? What simplifications are made and why are they acceptable here but not recommended in standard spectral clustering?
- It is assumed that K-means finds the global minimum.

Assumptions and consistency result

Here will denote by $\bar{L}, \bar{D}, \ldots$ the values of the quantities in the algorithm obtained from the expected Y matrix (with $E[Y] = \bar{[Y]} = [P_{ij}]$).

- Eigenvalue gap[1] $n^{-1/2} \log^2 n = \mathcal{O}(\lambda_K^2)$
- Node degrees are not too small $\tau_n = \frac{\min_i \bar{d}_i}{n} > \frac{2}{\ln n}$

- Denote by n_1 the size of the largest (true) cluster
- Let $M = \{i \in V, ||X_{i:} - \bar{X}_{i:}O||_2 \geq 1/\sqrt{2n_1}\}$; this defines the set of (possibly) misclustered points. $X_{i:}$ is the row of i in X, $\bar{X}_{i:}$ is its row in the X obtained from the expected Y, and O is an orthogonal matrix that "aligns" the two sets of columns.
- With the above assumptions, it can be proved that, with high probability,

$$|M| \;=\; o\left(\frac{n_1 \ln^2 n}{n\lambda_K^4 \tau_n^4}\right) \tag{3}$$

Intuition: why it should work

- First idea: the "expected" case is simple Note that $\bar{L}$ has constant rows in each cluster. Hence, its rank is equal to the number of clusters in the data, i.e. K. The number of non-zero eigenvalues (of $\bar{L}$) indicates this number.

  Also, since $\bar{L}$ has constant rows in each cluster, its K principal eigenvectors $\bar{X}$ will be piecewise-constant, i.e the rows of $\bar{X}$ will be the same whithin a cluster.

[1] Is this $\bar{\lambda_K}$?

- Second idea: how to show the actual X is "near" the "expected"

  This is a brilliant idea! First to see why a brilliant idea is needed, on a very simple example.

  Example Let $Y_{ij} \sim Bernoulli(\frac{1}{2})$ (iid) for all $i, j \in V$. Then,

$$\frac{1}{n^2}||Y - E[Y]||_F^2 \;=\; \frac{1}{4} \tag{4}$$

  The Frobenius norm sums over all entries in Y, and the normalization by n^2 computes the average. So, the matrix Y does not concentrate around its mean when $n \to \infty$! It will be hard to prove that its eigenvectors do! This is the hurdle. (The Laplacian will behave similarly, up to some scaling factors.)

  Now, the brilliant idea is to square Y (in the proof, L is squared). This does not chage the eigenvectors; if we can prove that Y^2 concentrates around $(E[Y])^2$, then we have a proof for the eigenvectors consistency (after calling in some Matrix Perturbation Theory).

$$(Y^2)_{ij} \;=\; \sum_k Y_{ik}Y_{kj} \;\sim\; Binom(\frac{1}{4}, n), \text{ for } i \neq j, \text{ and } \;\sim\; Binom(\frac{1}{2}, n), \text{ for } i = j \tag{5}$$

$$(E[Y])^2 \;=\; \frac{n}{4}\mathbf{1}_{n\times n} \tag{6}$$

  To make the story short, Y^2/n concentrates, and

$$\frac{1}{n^2}||Y^2/n - (E[Y])^2/n||_F^2 \;=\; o\left(\frac{\ln^2 n}{n}\right) \tag{7}$$

  (this bound is not trivial to prove).

3 The probabilistic method in graph theory – an example

Take an Erdos-Renyi graph on n nodes, with parameter $p \in (0, 1)$. What should be p so as to guarantee that the graph has no isolated nodes (w.h.p)? It is easy to see that p cannot be constant, it must grow with n. The question is how? We will prove that the correct rate is $p(n) = \frac{\ln n}{n}$. This function $p(n)$ is called a **threshold** because whenever $p > p(n)$ the graph has (asymptotically) almost surely no isolated nodes, and whenever $p < p(n)$ the graph will, again (asymptotically) almost surely have some.

Proof (after D. West) Set $X_i = 1$ if $i \in V$ is isolated. Let $X = \sum_i X_i$. Obviously, $E[X_i] = (1-p)^{n-1}$, $E[X] = n(1-p)^{n-1}$, and the graph has no isolated nodes iff $X = 0$.

Assume $p = c\frac{\ln n}{n}$, with $c > 1$. Let's make some (rather crude) approximations (assuming n large):

$$(1-p)^n = e^{n\ln(1-p)} = e^{n[-p+p^2/2-\ldots]} \approx e^{-np}, \quad (8)$$

$$E[X] \approx ne^{-np}/(1-p) = \approx n^{1-c} \quad (\text{since } 1-p \to 1), \quad (9)$$

and since $c > 1$ we have that $E[X] \to 0$. Since $X \in \{0, 1, \ldots\}$, it implies that $P[X = 0] \to 1$.

For the reverse relationship, we assume $c < 1$ and we need to use another proof technique. Exercise Fill in the gaps in this proof sketch By Markov's inequality, for any non-negative random variable

$$P[X = 0] \leq \frac{E[X^2]}{(E[X])^2} - 1 \quad (10)$$

Hence, we prove (with more crude approximations) that $E[X^2]$ approaches $(E[X])^2$. Hence $P[X = 0] \to 0$ and the graph has isolated nodes w.h.p.

A remarkable fact in random graph theory is that, at exactly the same threshold, the graph also becomes connected.

4 Homomorphisms and graph convergence

(After Lovasz, "Large networks and graph limits") Let $\mathcal{G} = (V, E)$, $\mathcal{H} = (U, F)$ be two graphs. A **homomorphism** between $\mathcal{G}$ and $\mathcal{H}$ is a function $\phi : V \to U$ with the property that if $ij \in E$, then $\phi(i)\phi(j) \in F$. In other words, if i, j are connected by an edge in $\mathcal{G}$, their images are also connected in $\mathcal{H}$. In general, a homomorphism need not be injective, and non-edges in $\mathcal{G}$ can map to edges in $\mathcal{H}$. In the theory of *graph limits*, the *number of homomorphisms* between $\mathcal{G}$ and $\mathcal{H}$ denoted $\hom(\mathcal{G}, \mathcal{H})$ plays a central role.

It is helpful to consider, for a fixed graph $\mathcal{G}$, the infinite vector $h(\mathcal{G}) = [\ldots \hom(\mathcal{H}, \mathcal{G}) \ldots]$ where $\mathcal{H}$ ranges over all graphs. Two graphs $\mathcal{G}, \mathcal{G}'$ may be compared by comparing their vectors $h(\mathcal{G}), h(\mathcal{G}')$. Interestingly, $h(\mathcal{G}) = h(\mathcal{G}')$ iff $\mathcal{G}$ and $\mathcal{G}'$ are isomorphic (which means if they are the same graph, up to a renaming of the nodes) (Theorem 5.29). Moreover, the graph $\mathcal{G}$ is uniquely determined by the finite set of graphs $\mathcal{H}$ with no more nodes than $\mathcal{G}$. The same is true for $\hom(\mathcal{G}, \ldots)$.

Homomorphism densities are normalized number of homomorphisms. Let $|V| = n$, $|U| = k$.

$$t(\mathcal{H}, \mathcal{G}) = \frac{\hom(\mathcal{H}, \mathcal{G})}{n^k} \quad (11)$$

is the probability that a random subset of k nodes from $\mathcal{G}$ (with replacement) induces a graph homomorphic with $\mathcal{H}$.

For sparse graphs, the homomorphism densities become too small to be useful, and one works with **homomorphism frequencies**, defined by

$$t^*(\mathcal{H}, \mathcal{G}) \;=\; \frac{\hom(\mathcal{H}, \mathcal{G})}{n} \tag{12}$$

The interpretation of t^* is as follows: Label one node of U with label 1. Denote by $\hom_i(\mathcal{H}, \mathcal{G})$ the number of homomorphisms of $\mathcal{H}$ which map node 1 to $i \in V$. Since the graph is finite degree, if $\mathcal{H}$ is connected, all the homomorphisms will be in a k-neighborhood of node i, which contains only a bounded number of nodes, no matter what n is. Hence $\hom_i(\mathcal{H}, \mathcal{G})$ is bounded for all i. Now $t^*(\mathcal{H}, \mathcal{G})$ is the average over $i \in V$ of $\hom_i(\mathcal{H}, \mathcal{G})$.

In the same way as for the unnormalized $\hom(\ldots, \mathcal{G})$, we can ask if the vector of all $t(\mathcal{H}, \mathcal{G})$ determines $\mathcal{G}$. The answer is that it "almost" determines it, up to a "blow-up"[2].

By representing a graph as a vector $t(\mathcal{G})$ we can now consider when a sequence of graphs converges. It is understood that in a sequence of graphs, the number of nodes n tends to infinity, so we can assume w.l.o.g that $\mathcal{G}_n$ has n nodes. We say that the sequence $\{\mathcal{G}_n\}_n$ is **convergent** iff $t(\mathcal{H}, \mathcal{G}_n) \to t(\mathcal{H})$ for every $\mathcal{H}$.

The limit object is not a graph, but a **graphon** (described next). Note that this is not the only way to define a graph limit, but it is one in which graph properties are preserved.

5 Graphons

A **graphon** is a function $f : [0,1]^2 \to [0, \infty)$ which is measurable, bounded, symmetric (i.e. $f(x,y) = f(y,x)$). Intuitively, the interval $[0,1]$ represents the vertex set V, and $f(x,y)$ represents the probability of edge (x,y).

Hence, the degree (distribution) of node x defined by the marginal

$$d_f(x) \;=\; \int_0^1 f(x,y)dy \tag{13}$$

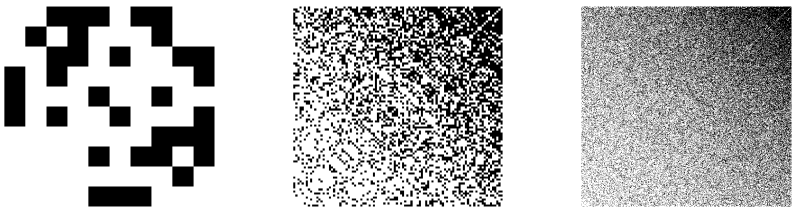
corresponds to $\lim_{i/n \to x} \frac{d_i}{n}$.

[2] A "blow-up" is an operation in which each node is replicated p times, along with its neighborhood. The normalization in 11 makes the blown-up graph indistinguishable from the original graph.

As permuting the numbering of the nodes of a graph leaves the graph invariant, a graphon f is defined up to "permutations" of $[0,1]$, called measure preserving bijections.

The example below shows the limit of the random graph sequence given by $P_{ij}^{(n)} = e^{-(i+j)/n}$ to the graphon $f(x,y) = e^{-(x+y)}$.

$n = 10$ $n = 100$ $n = 500$



Limits of sparse graphs are defined in terms of the homomorphism frequencies t^*, and are called **graphimes**.

6 SBM and graphon estimation

One can approximate a density f over $[0,1]^2$, arbitrarily closely by a piecewise constant function over a grid of intervals that covers $[0,1]^2$. Such an approximation $\hat{f}$ is nothing else than a Stochastic Block Model. Therefore, estimation of SBM's and of graph limits are deeply related[3]

A generative model

Let f be a graphon, and (ρ_n) an unbounded sequence of edge density parameters (without it, the graphs generated would have bounded degree – too sparse).

Given n

1. Sample nodes $\xi_{1:n}$ uniformly i.i.d from $[0,1]$
2. Sample edges Y_{ij}, $1 \leq i < j \leq n$ from $Bernoulli(P_{ij})$ with

$$P_{ij} \;=\; \rho_n f(\xi_i, \xi_j) \tag{14}$$

[3]The relationship follows from Szemeredy's Regularity Lemma which states that any graph can be approximated arbitrarily closely by a SBM with a large enough number of blocks. In general, "large enough" is a very large number (a tower of powers). So, to "learn" a graph (limit) we must assume that it is "nice", perhaps in the sense that it is described by not so many blocks.

An idealized graphon estimation algorithm

This algorithm outputs a partition of the nodes into K clusters, and, for each pair of labels , $(k, k') \in \{1, \dots K\}^2$ a value $\hat{f}_{k,k'}$. The set of values $\{\hat{f}_{kk'}, (k, k') \in \{1, \dots K\}^2\}$ represent a graphon $\hat{f}$ that is piecewise constant over the edges with ends $z_k, z_{k'}$. The node locations $\xi_{1:n}$ are not determined, because graphons are defined up to bijective measure preserving transformations of $[0, 1]$. We denote by z_i the cluster label of point i.

Given a graph with edgeset Y and K the "number of blocks"

- For all possible assignments $z_{1:n}$ of nodes into blocks
  1. Estimate the SBM matrix $B(z)$

$$B_{kk}(z) \;=\; \frac{\sum_{z_i=z_j=k} Y_{ij}}{n_k(n_k-1)} \quad B_{kk'}(z) \;=\; \frac{\sum_{z_i=k,z_j=k'} Y_{ij}}{n_k n_{k'}}, \text{ for } k \neq k' \tag{15}$$

  2. Estimte its log-likelihood $l(z)$
- Select the $\hat{B} = B(\hat{z})$ with $\hat{z} = \text{argmax}_z\, l(z)$

Theorem When does it work?
Assumptions

- smoothness f is Hölder continuous with parameter $\alpha \in (0, 1]$

$$\frac{|f(x,y) - f(x',y')|}{||(x,y) - (x',y')||^\alpha} \;\leq\; M \;\leq\; \infty \tag{16}$$

- f not too sparse $\inf f \geq \varepsilon > 0$
- scaling ρ_n growing not too slowly $\rho_n = \omega(n^{-1} \log^3 n)$[4]
- K grows unbounded with n
- partition not too far from uniform Let n_k the size of C_k, $n_{min} = \min_k n_k$, $\bar{n} = n/K$
- sufficient samples (edges) in each grid cell $C_k \times C_{k'}$ $n_{min}^2 \rho_n = \omega(\log n)$, $\bar{n}^2 \rho_n = \omega(K^2 \log^3 n)$

[4] $f_n = \omega(g_n)$ means $g_n = o(f_n)$, i.e the rate of f_n is strictly larger than the growth rate of g_n.

It can be shown, under these conditions that the non-parametric estimator of f given by $(\hat{z}, \hat{B})$ is consistent, i.e. that it converges to the true graphon f when $n \to \infty$.

The convergence is in mean squared error

$$\inf_{\sigma} \int \int_{[0,1]^2} |f(\sigma(x), \sigma(y)) - \hat{f}(x, y)|^2 dx dy \tag{17}$$

where σ ranges over all measure preserving bijections (that is "permutations" of $[0, 1]$) of $[0, 1]$ into itself.