Probability on Graphs: Techniques and Applications to Data Science

0 - Preliminaries

Sébastien Roch
UW–Madison
Mathematics

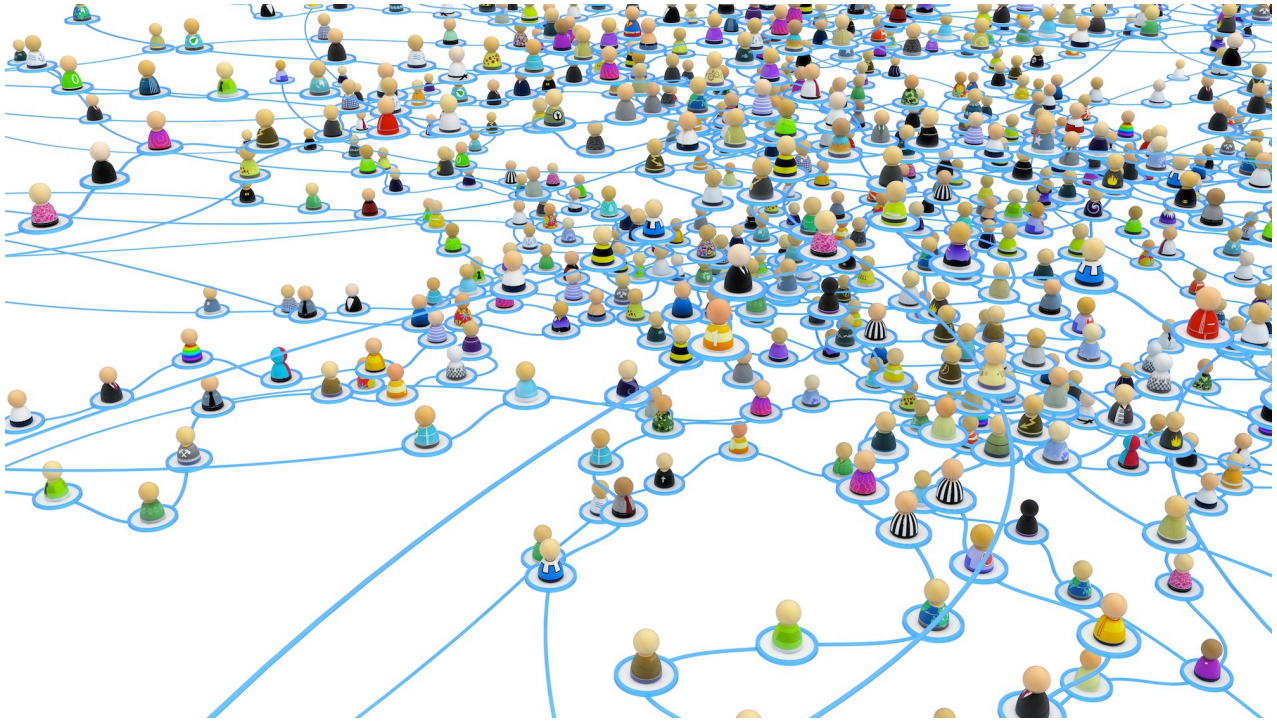
July 25, 2018

Go deeper

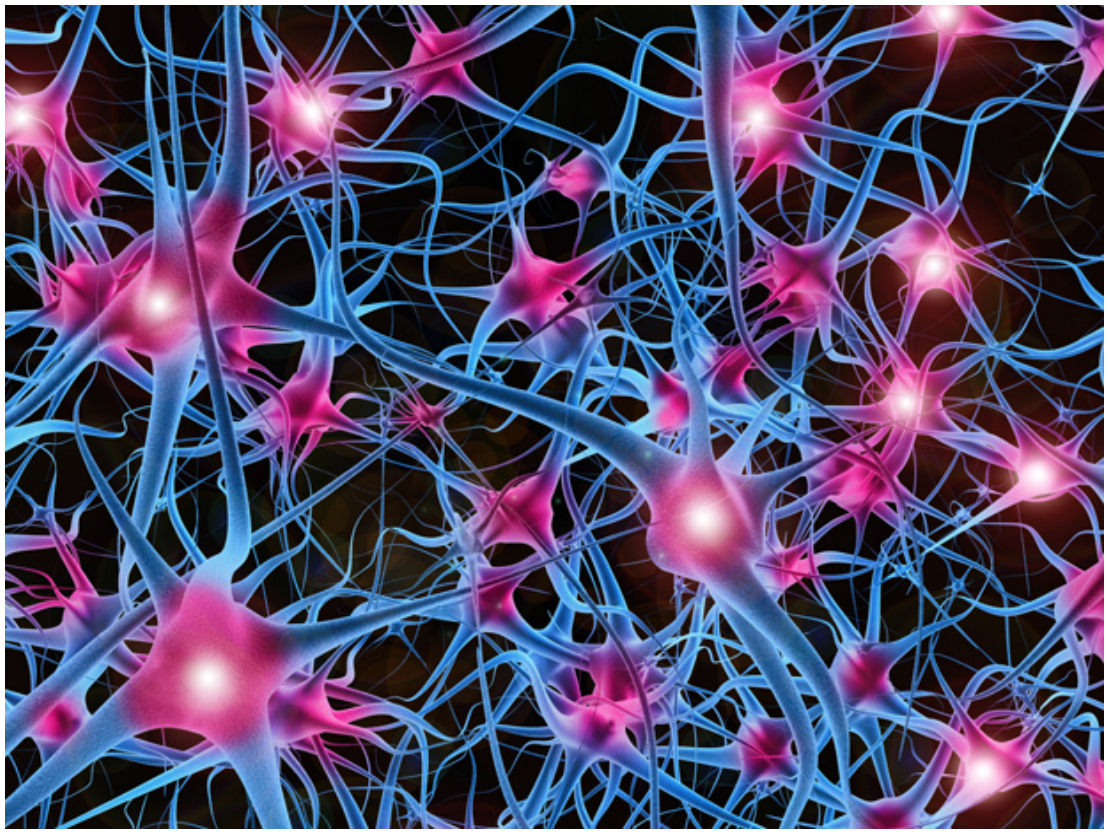
A lot more details and examples in the lecture notes at:

`http://www.math.wisc.edu/~roch/mdp/`

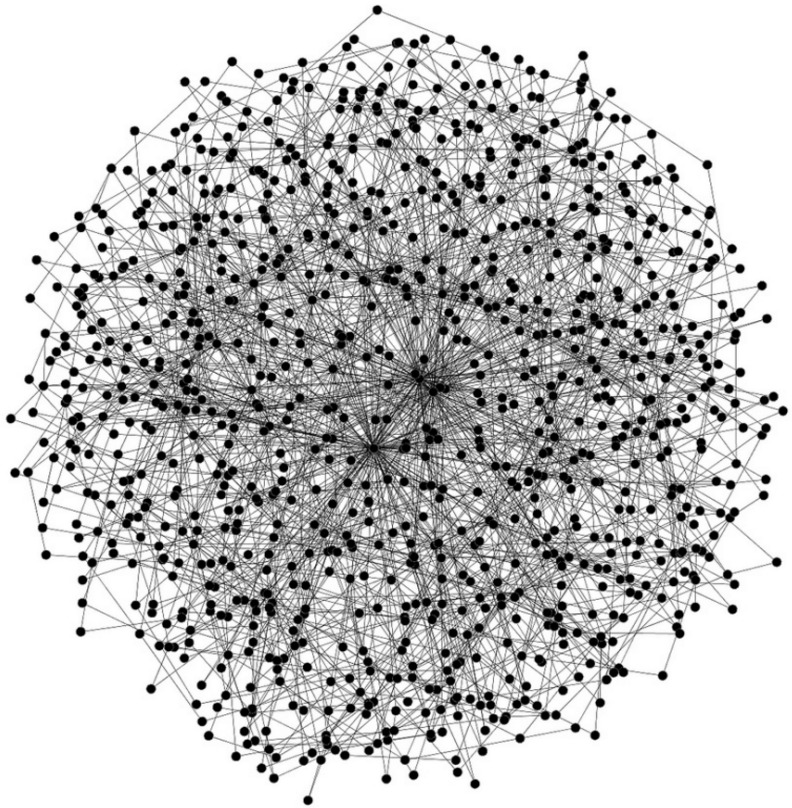
Networks are ubiquitous: Social networks



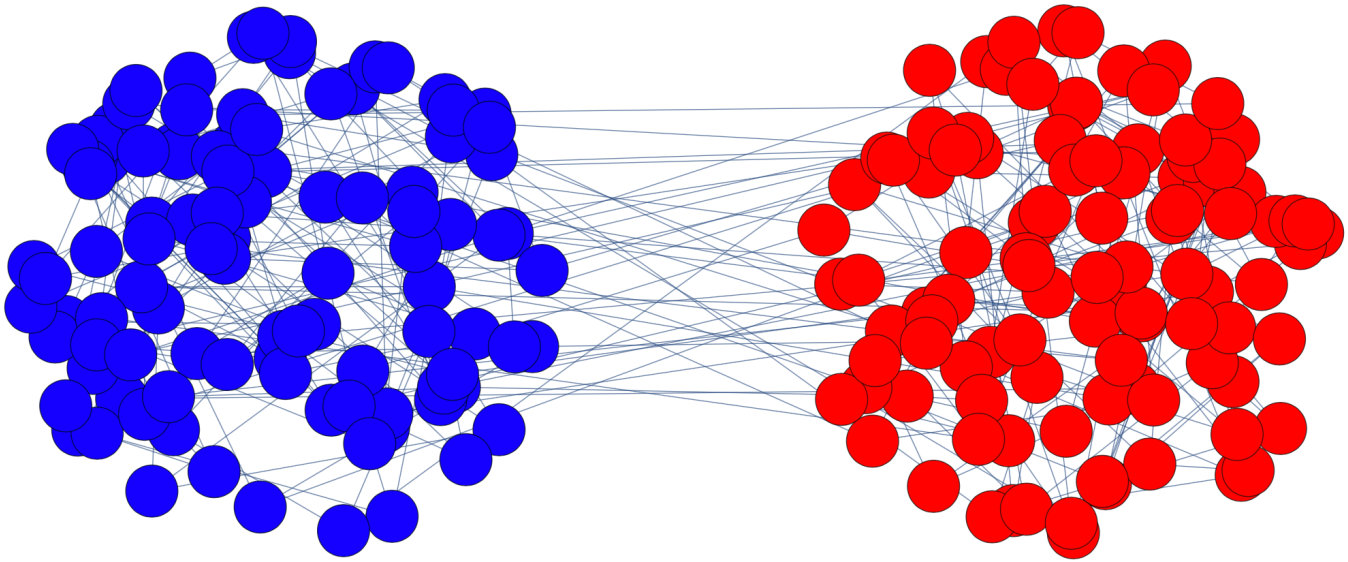
Networks are ubiquitous: Biological networks



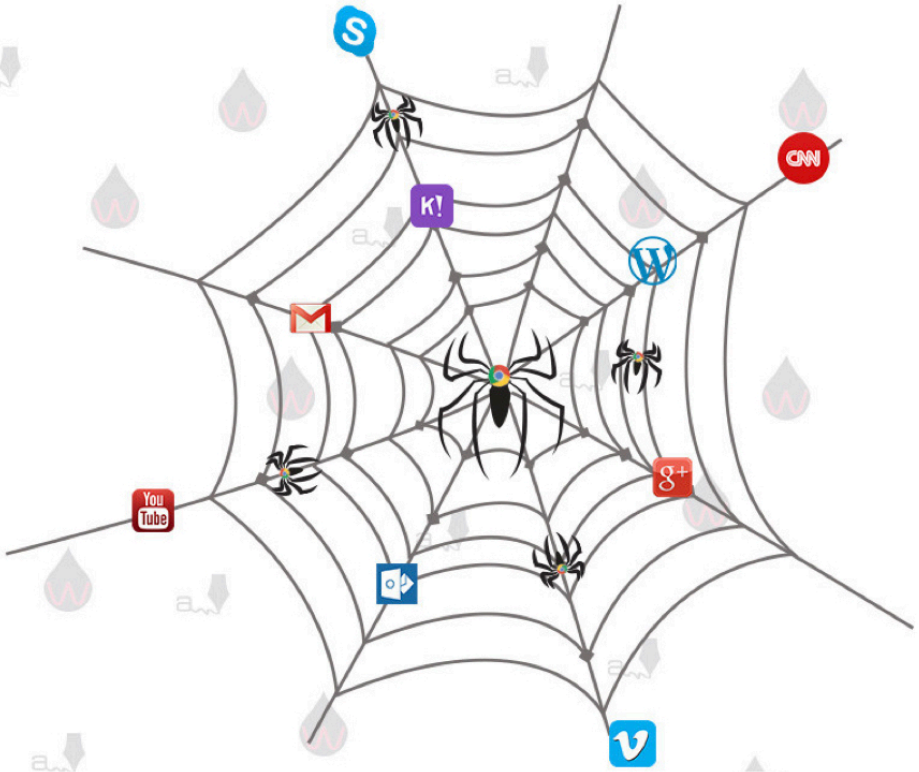
Data science: Network modeling



Data science: Network processes



Data science: Network sampling



Graphs

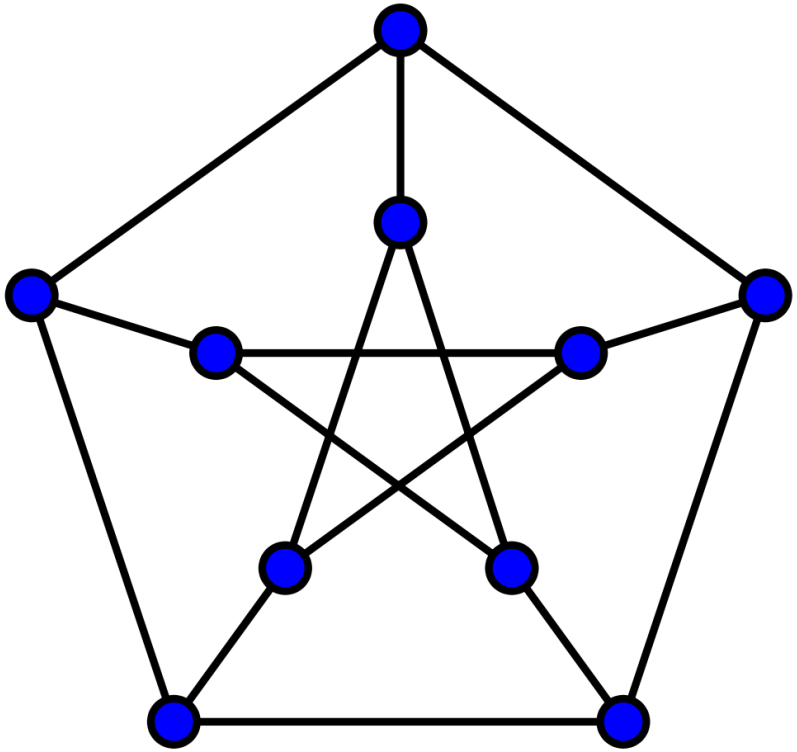
Definition

An *(undirected) graph* is a pair $G = (V, E)$ where V is the set of *vertices* and

$$E \subseteq \{\{u, v\} : u, v \in V\},$$

is the set of *edges*.

An example: the Petersen graph



Basic definitions

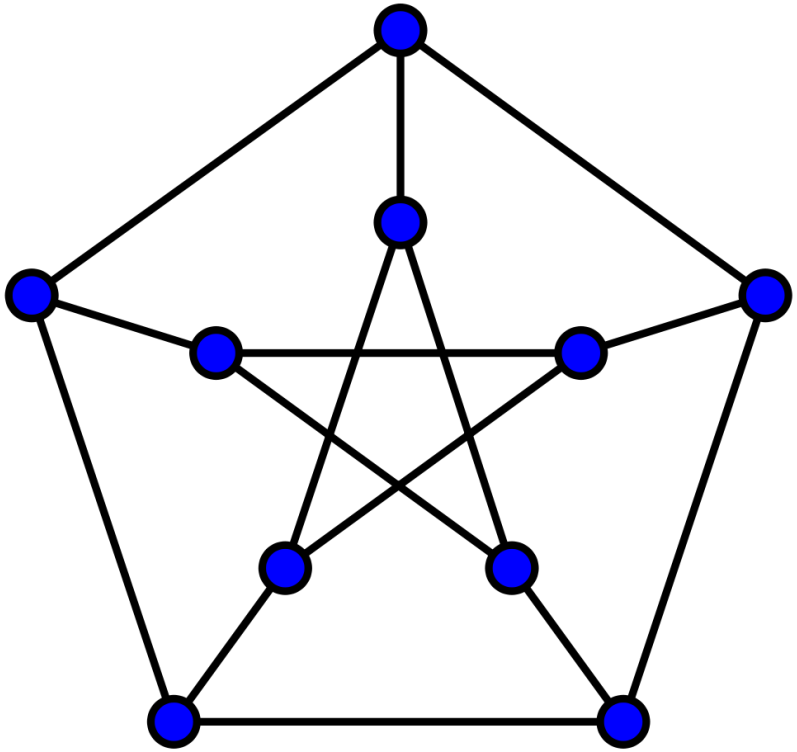
Definition (Neighborhood)

Two vertices $u, v \in V$ are *adjacent*, denoted by $u \sim v$, if $\{u, v\} \in E$. The set of adjacent vertices of v, denoted by $N(v)$, is called the *neighborhood* of v and its size, i.e. $\delta(v) := |N(v)|$, is the *degree* of v. A vertex v with $\delta(v) = 0$ is called *isolated*.

Example

All vertices in the Petersen graph have degree 3. In particular there is no isolated vertex.

An example: the Petersen graph



Paths and connectivity

Definition (Paths)

A *path* in G is a sequence of vertices $x_0 \sim x_1 \sim \cdots \sim x_k$. The number of edges, k, is called the *length* of the path. If $x_0 = x_k$, we call it a *cycle*. We write $u \leftrightarrow v$ if there is a path between u and v. The equivalence classes of $\leftrightarrow$ are called *connected components*. The length of the shortest path between two vertices u, v is their *graph distance*, denoted $d_G(u, v)$.

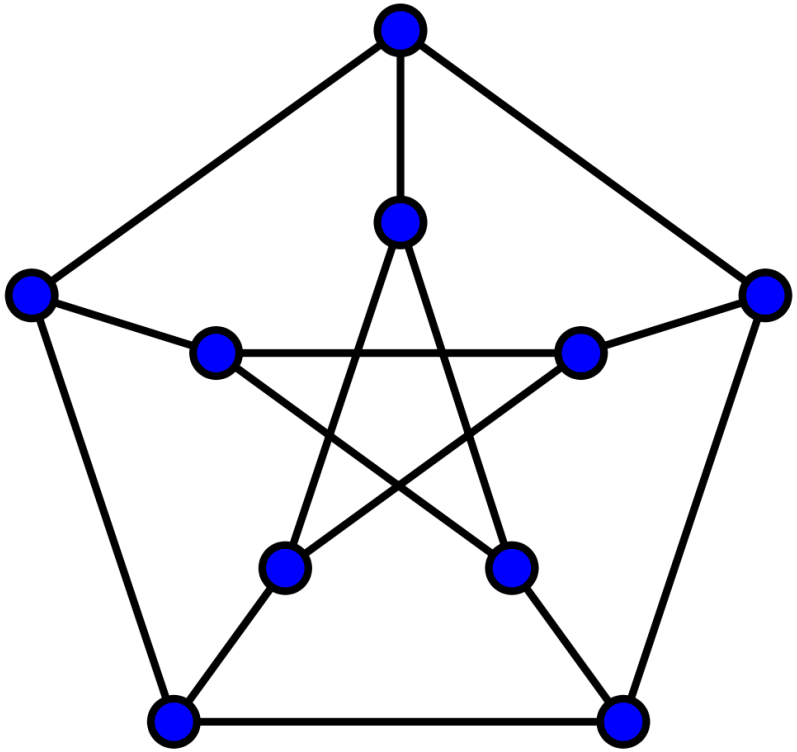
Definition (Connectivity)

A graph is *connected* if any two vertices are linked by a path, i.e., if $u \leftrightarrow v$ for all $u, v \in V$.

Example

The Petersen graph is connected.

An example: the Petersen graph



Adjacency matrix

Definition

Let $G = (V, E)$ be a graph with $n = |V|$. The *adjacency matrix* A of G is the $n \times n$ matrix defined as $A_{xy} = 1$ if $\{x, y\} \in E$ and 0 otherwise.

Example

The adjacency matrix of a *triangle* (i.e. 3 vertices with all edges) is

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

Examples of finite graphs

- K_n: clique with n vertices, i.e., graph with all edges present
- C_n: cycle with n non-repeated vertices
- $\mathbb{H}^n$: n-dimensional hypercube, i.e., $V = \{0,1\}^n$ and $u \sim v$ if u and v differ at one coordinate

Erdös-Rényi random graph

Definition

Let $V = [n]$ and $p \in [0,1]$. The *Erdös-Rényi graph* $G = (V, E)$ on n vertices with density p is defined as follows: for each pair $x \neq y$ in V, the edge $\{x, y\}$ is in E with probability p independently of all other edges. We write $G \sim \mathbb{G}_{n,p}$ and we denote the corresponding measure by $\mathbb{P}_{n,p}$.

Questions:

- What is the probability of observing a triangle?
- Is G connected?
- What is the typical chromatic number (i.e., the smallest number of colors needed to color the vertices so that no two adjacent vertices share the same color)?

Other random graph models

- Preferential attachement
- Small world
- Fixed degree distribution

Random walk on a network

Definition

Let $G = (V, E)$ be a graph. Let $c : E \to \mathbb{R}_+$ be a positive edge weight function on G. We call $\mathcal{N} = (G, c)$ a *network*. Random walk on $\mathcal{N}$ is the Markov chain on V, started at an arbitrary vertex, which at each time picks a neighbor of the current state proportionally to the weight of the corresponding edge.

Questions:

- How often does the walk return to its starting point?
- How long does it take to visit all vertices once or a particular subset of vertices for the first time?
- How fast does it approach stationarity?

Other sampling schemes

- Random walks with restarts
- Branching random walks
- Random sample of vertices and their neighbors

Undirected graphical models I

Definition

Let S be a finite set and let $G = (V, E)$ be a finite graph. Denote by $\mathcal{K}$ the set of all cliques of G. A positive probability measure μ on $\mathcal{X} := S^V$ is called a *Gibbs random field* if there exist *clique potentials* $\phi_K : S^K \to \mathbb{R}$, $K \in \mathcal{K}$, such that

$$\mu(x) = \frac{1}{\mathcal{Z}} \exp\left(\sum_{K \in \mathcal{K}} \phi_K(x_K)\right),$$

where x_K is x restricted to the vertices of K and $\mathcal{Z}$ is a normalizing constant.

Undirected graphical models II

Example

For $\beta > 0$, the *ferromagnetic Ising model* with inverse temperature β is the Gibbs random field with $S := \{-1, +1\}$, $\phi_{\{i,j\}}(\sigma_{\{i,j\}}) = \beta\sigma_i\sigma_j$ and $\phi_K \equiv 0$ if $|K| \neq 2$. The function $\mathcal{H}(\sigma) := -\sum_{\{i,j\}\in E} \sigma_i\sigma_j$ is known as the *Hamiltonian*. The normalizing constant $\mathcal{Z} := \mathcal{Z}(\beta)$ is called the *partition function*. The states $(\sigma_i)_{i\in V}$ are referred to as *spins*.

Questions:

- How fast is correlation decaying?
- How to sample efficiently?
- How to reconstruct the graph from samples?

Other graphical models

- Gaussian graphical models
- Bayes nets
- Latent graphical models

Go deeper

More details and examples on basic models at:

`http://www.math.wisc.edu/~roch/mdp/`

For more on probability on graphs in general, see e.g. (available online):

- *Probability on Graphs* by Grimmett
- *Probability on Trees and Networks* by Lyons with Peres

Probability on Graphs: Techniques and Applications to Data Science

1 - First and second moment methods

Sébastien Roch
UW–Madison
Mathematics

July 25, 2018

Markov's inequality

Theorem (Markov's inequality)

Let X be a non-negative random variable. Then, for all $b > 0$,

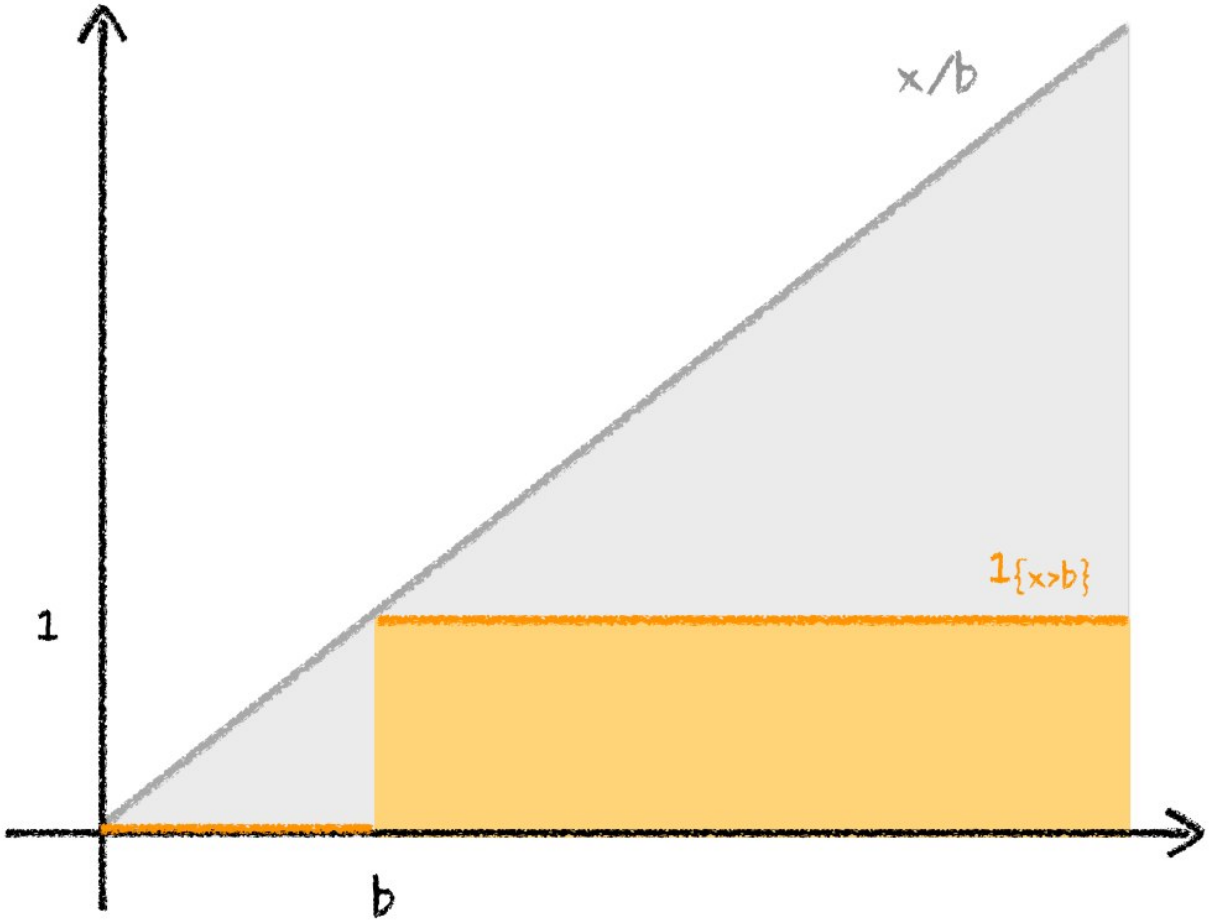
$$\mathbb{P}[X \geq b] \leq \frac{\mathbb{E}X}{b}.$$

Proof:

$$\mathbb{E}X \geq \mathbb{E}[X; X \geq b] \geq \mathbb{E}[b; X \geq b] = b\,\mathbb{P}[X \geq b].$$

■

Markov's inequality: Proof by picture



Chebyshev's inequality

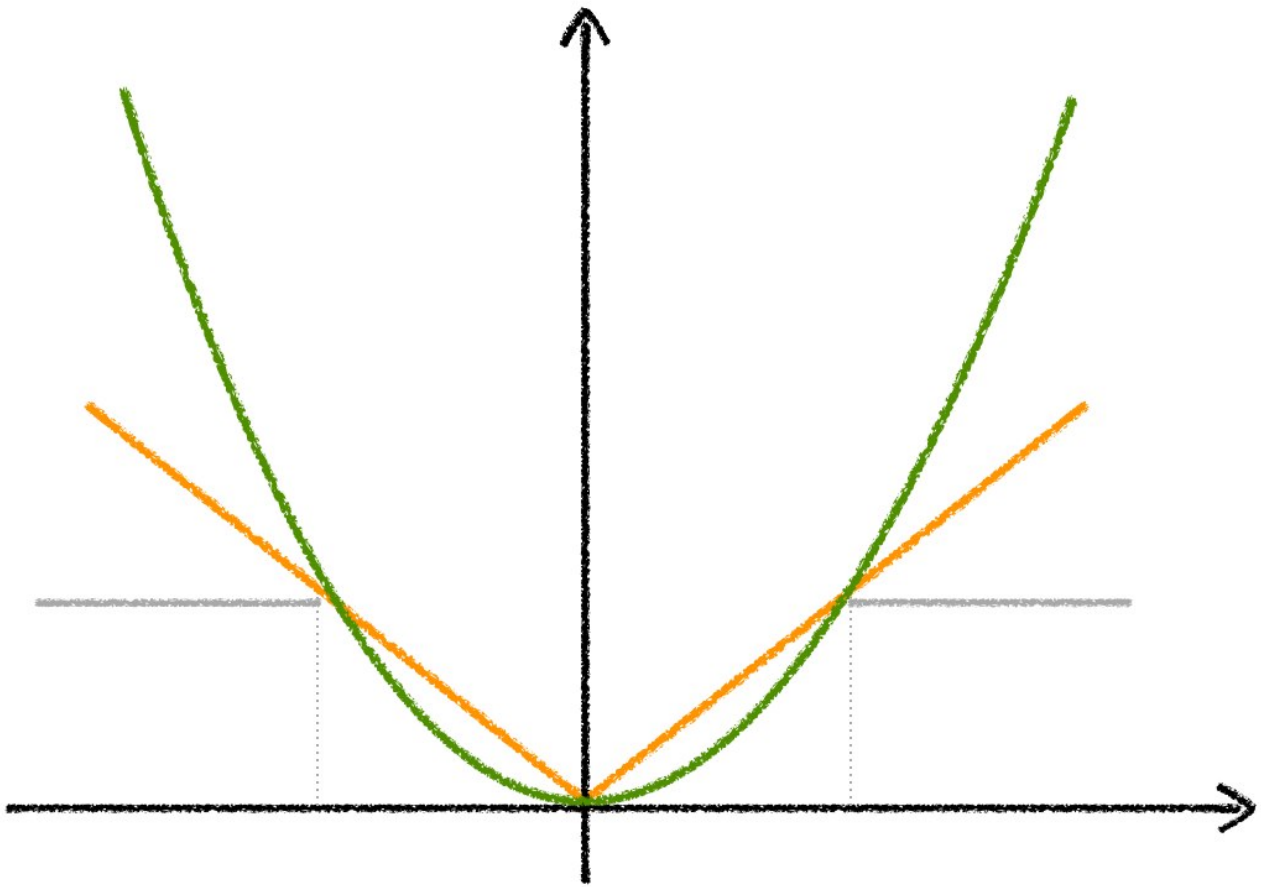
Theorem (Chebyshev's inequality)

Let X be a random variable with $\mathbb{E}X^2 < +\infty$. Then, for all $\beta > 0$,

$$\mathbb{P}[|X - \mathbb{E}X| > \beta] \leq \frac{\mathrm{Var}[X]}{\beta^2}.$$

Proof: This follows immediately by applying Markov's inequality to $|X - \mathbb{E}X|^2$ with $b = \beta^2$. ■

Chebyshev's inequality: Proof by picture



First moment method

Theorem (First moment method)

If X is a non-negative, integer-valued random variable, then

$$\mathbb{P}[X > 0] \leq \mathbb{E}X.$$

Proof: Take $b = 1$ in Markov's inequality. ■

That is: if X has "small" expectation, then its value is 0 with "large" probability. Typically used in the following way: one wants to show that a "bad event" does not occur with high probability; the random variable X counts the number of such "bad events." In that case, X is a sum of indicators and the theorem reduces to the *union bound*.

Going in the other direction

The first moment method gives an *upper bound* on the probability that a non-negative, integer-valued random variable is positive—provided its expectation is small. Suppose we want a *lower bound*. Note that a large expectation does not suffice.

Example

Say X_n is n^2 with probability $1/n$, and 0 otherwise. Then $\mathbb{E}X_n = n \to +\infty$, yet $\mathbb{P}[X_n > 0] \to 0$.

Second moment method

Theorem (Second moment method)

If X is a non-negative, integer-valued random variable, then

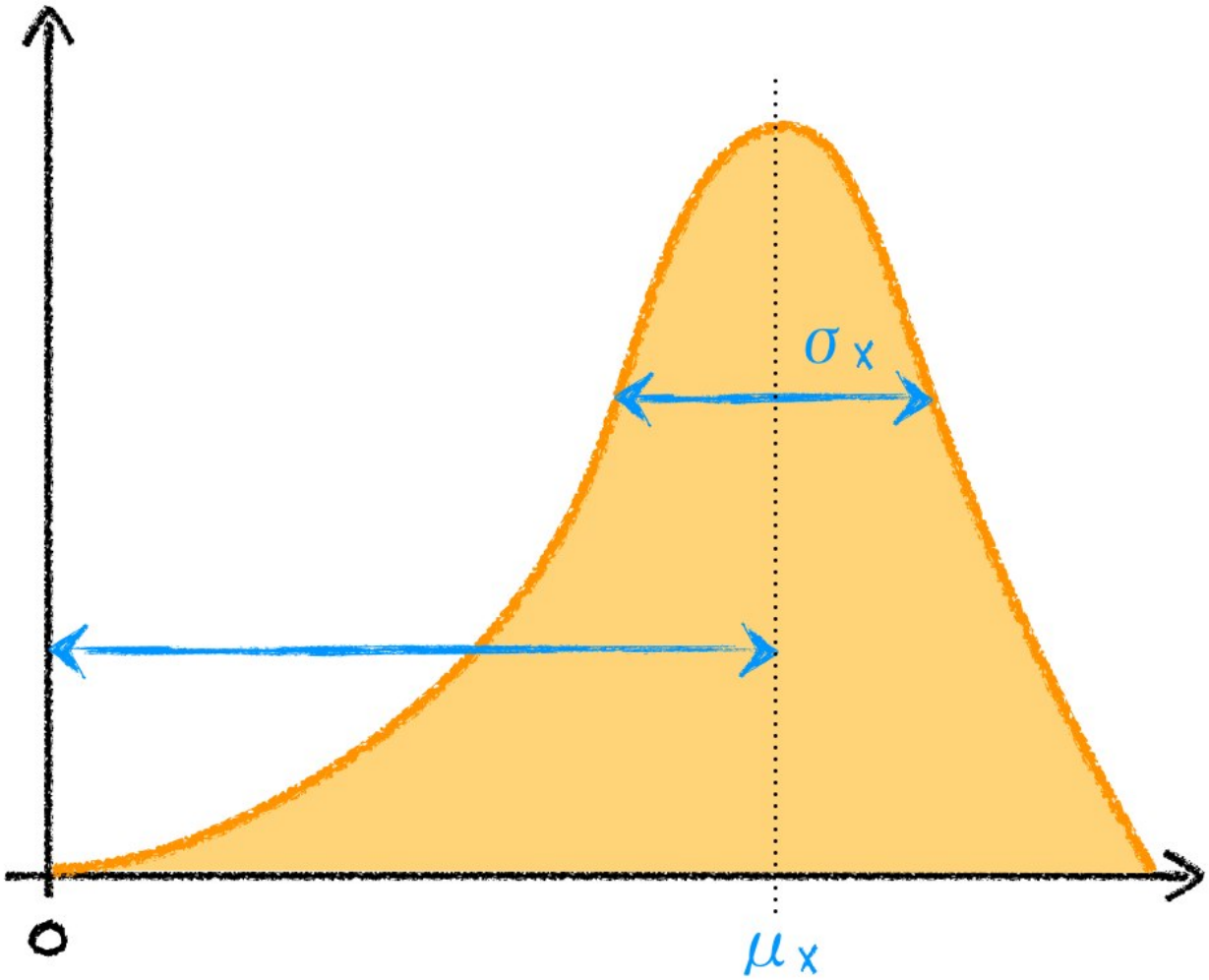
$$\mathbb{P}[X > 0] \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} \left(= 1 - \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2 + \mathrm{Var}[X]} \right).$$

Proof (of weaker version): By Chebyshev's inequality,

$$\mathbb{P}[X = 0] \leq \mathbb{P}[|X - \mathbb{E}X| \geq \mathbb{E}X] \leq \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2}.$$

■

Second moment method: Proof by picture



First and second moment methods: summary

If X is a non-negative, integer-valued random variable, then

$$\mathbb{P}[X > 0] \leq \mathbb{E}X,$$

and

$$\mathbb{P}[X > 0] \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}.$$

Threshold phenomena

Consider the Erdös-Rényi random graph. A *threshold function* for a graph property P is a function $r(n)$ such that

$$\lim_n \mathbb{P}_{n,p_n}[G_n \text{ has property } P] = \begin{cases} 0, & \text{if } p_n \ll r(n) \\ 1, & \text{if } p_n \gg r(n), \end{cases}$$

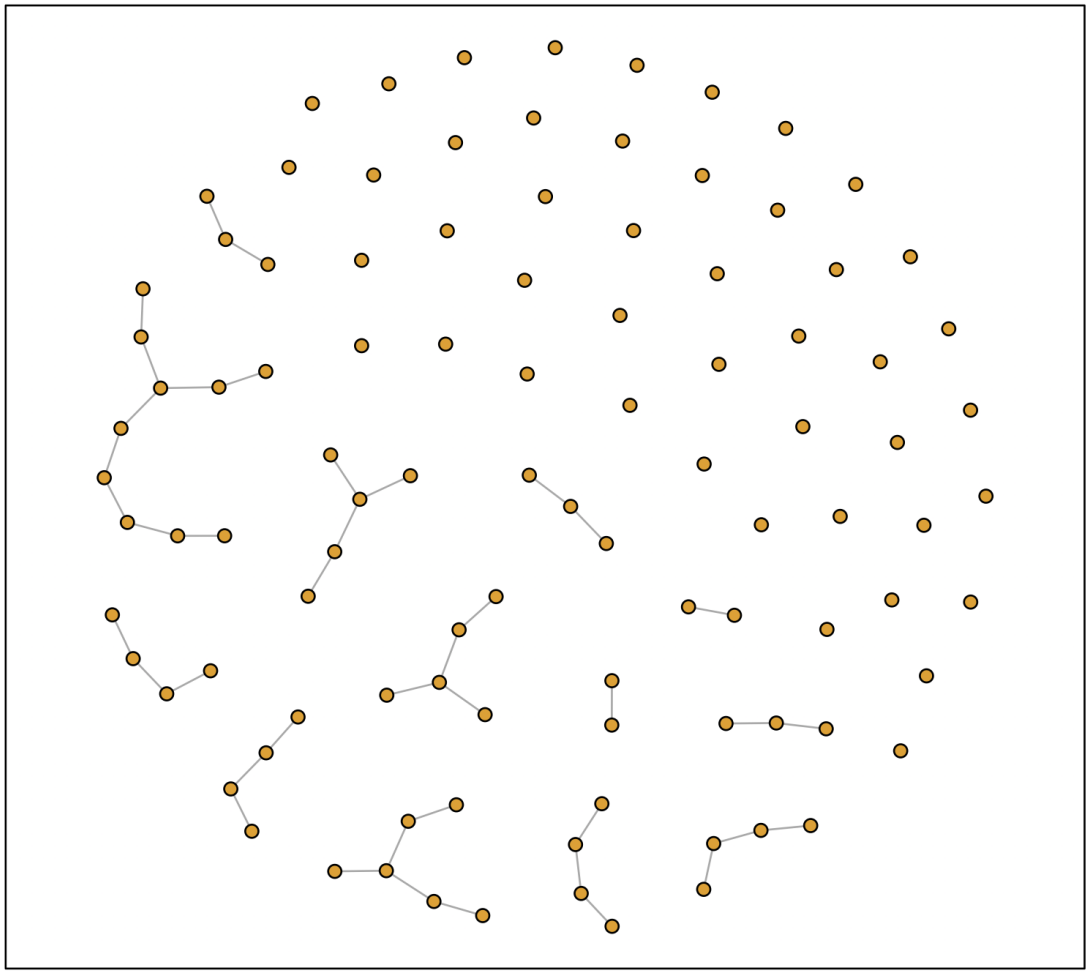
where $G_n \sim \mathbb{G}_{n,p_n}$ is an Erdös-Rényi graph with n vertices and density p_n.

Connectivity via isolated vertices

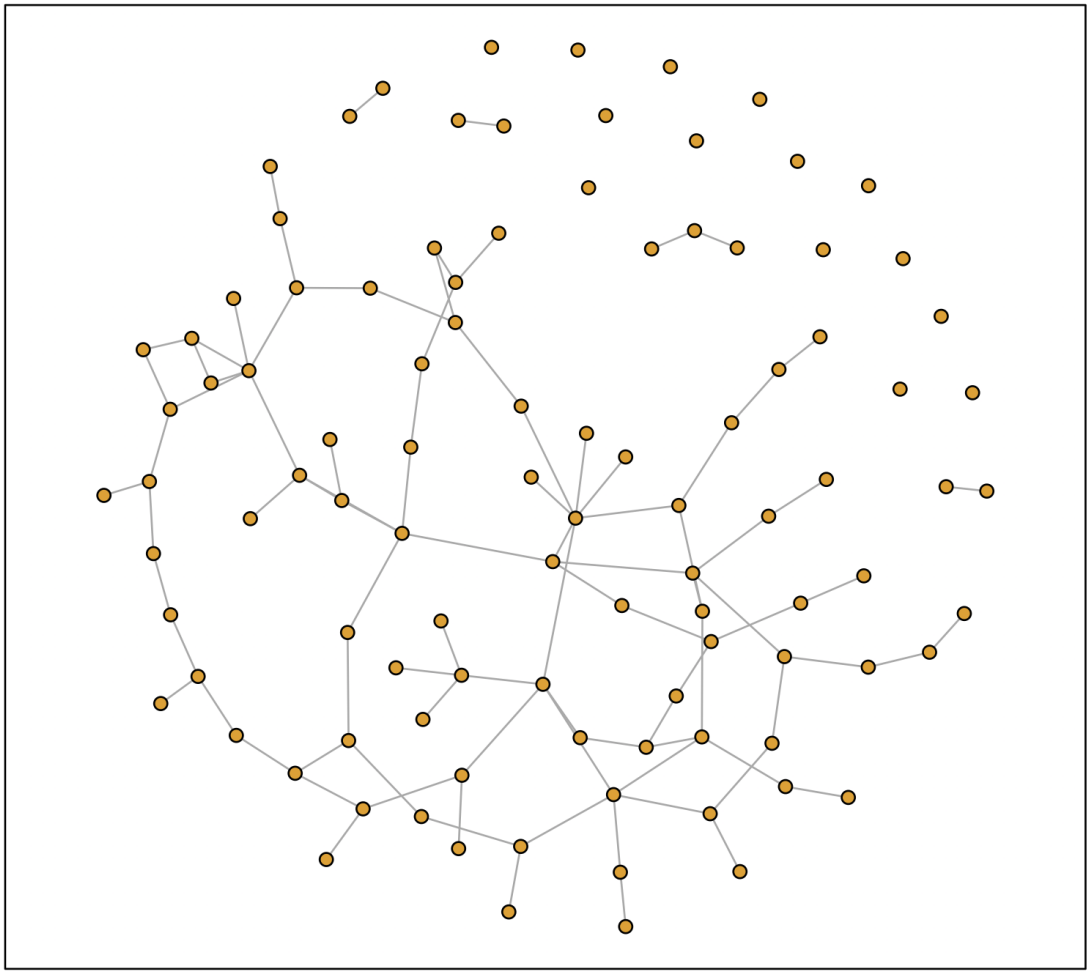
We use the first and second moment methods to show that the threshold function for connectivity in the Erdös-Rényi random graph is $\frac{\log n}{n}$.

We prove this result by deriving the threshold function for the presence of isolated vertices. Of course isolated vertices imply a disconnected graph. What is less obvious: the two thresholds actually *coincide*.

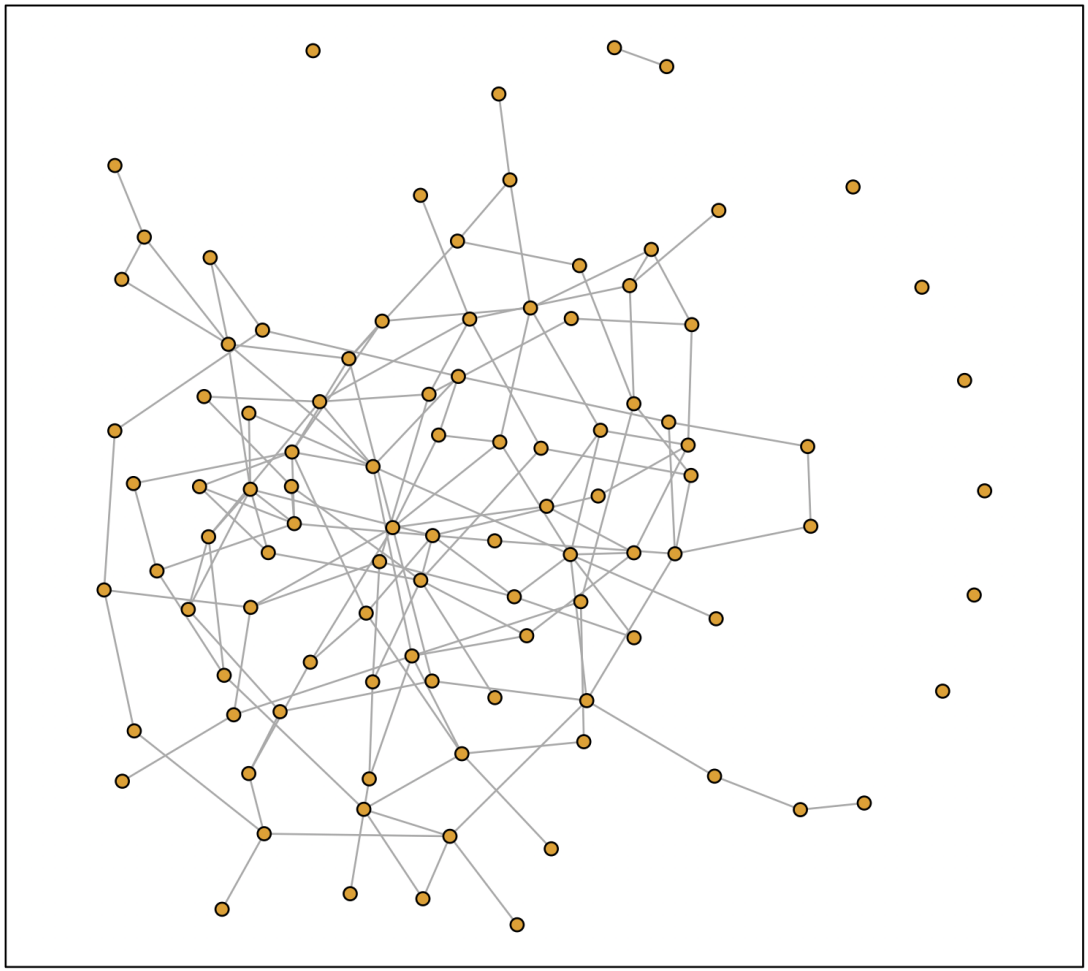
Erdös-Rényi with $n = 100$ and $p_n = 1/100$



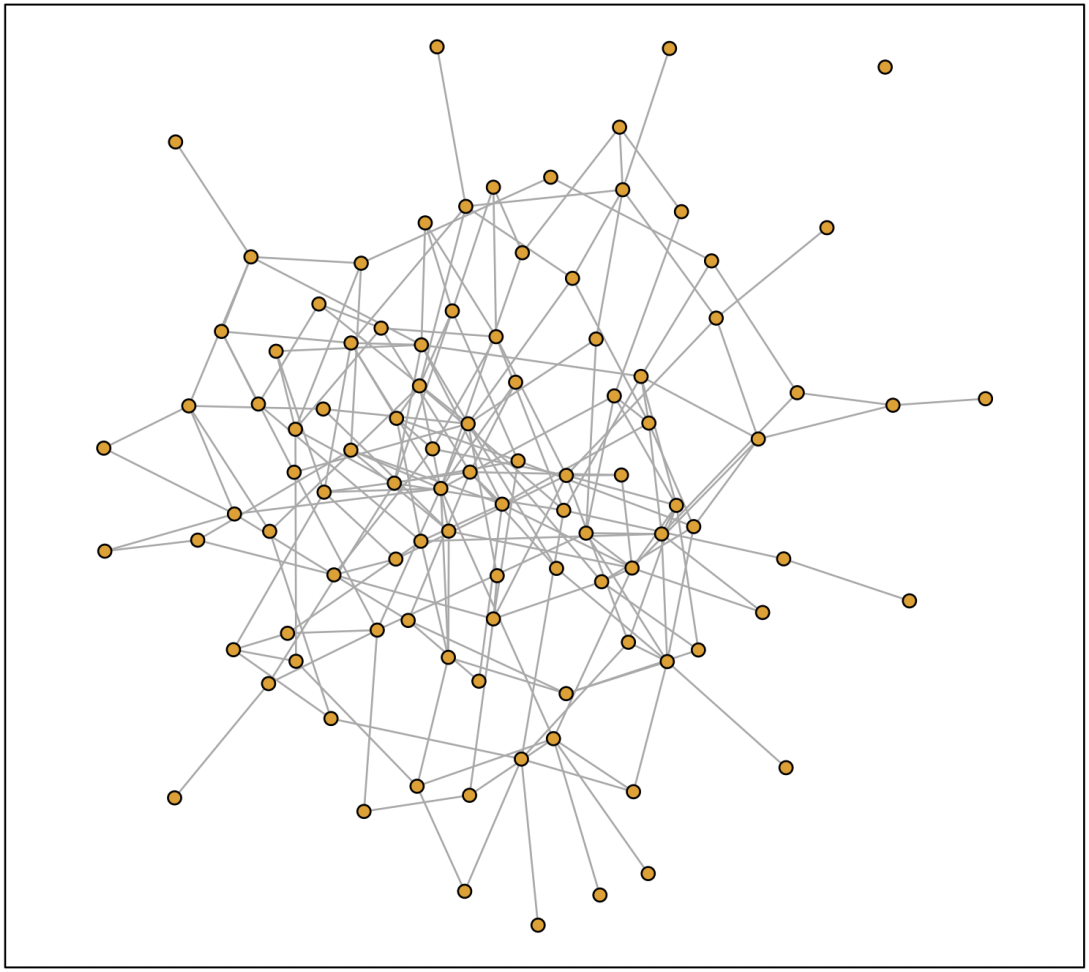
Erdös-Rényi with $n = 100$ and $p_n = 2/100$



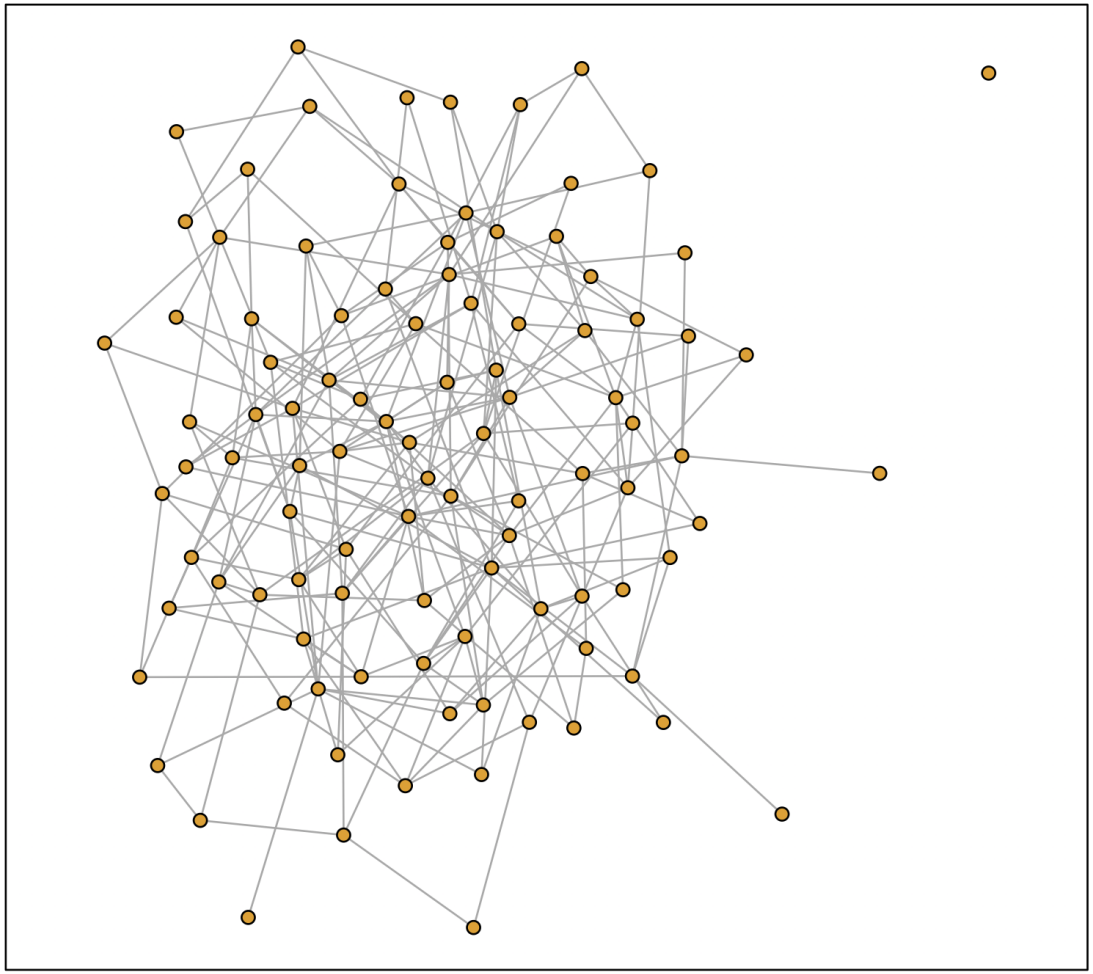
Erdös-Rényi with $n = 100$ and $p_n = 3/100$



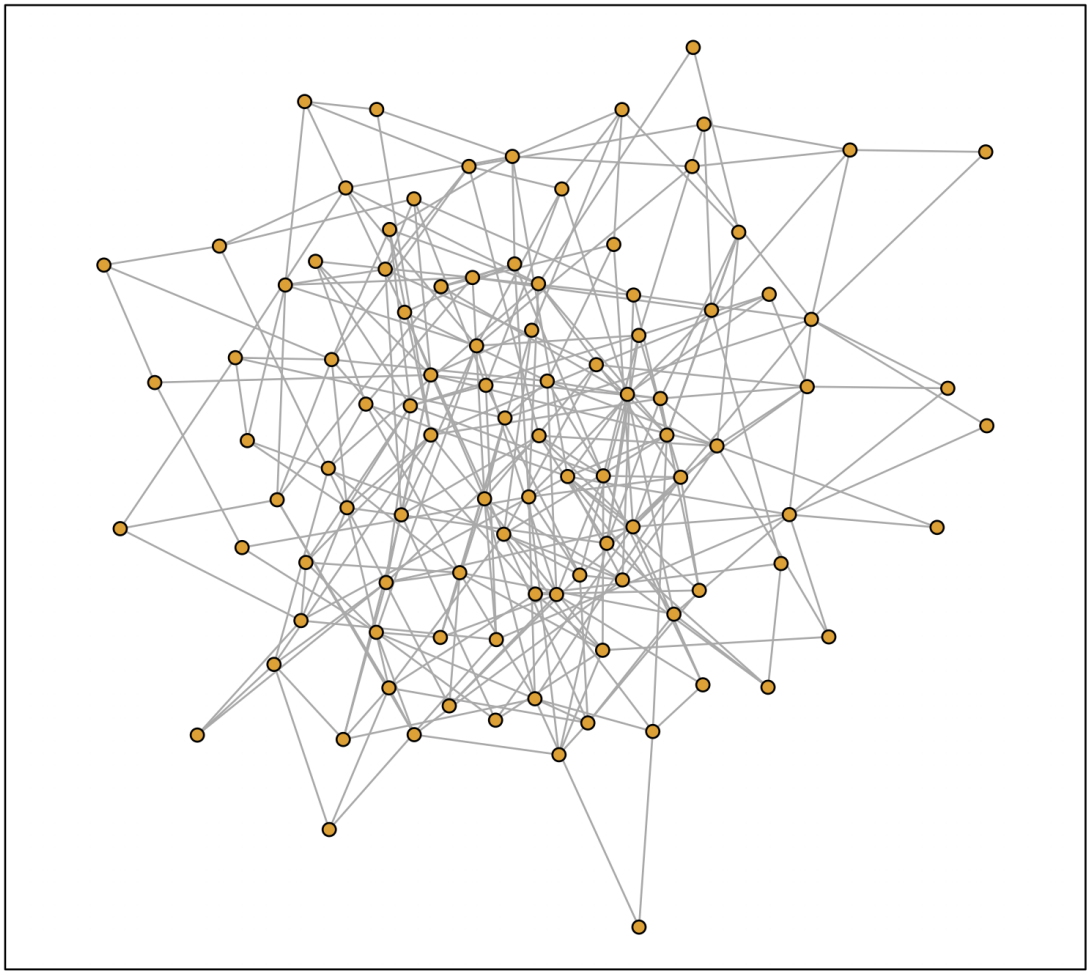
Erdös-Rényi with $n = 100$ and $p_n = 4/100$



Erdös-Rényi with $n = 100$ and $p_n = 5/100$



Erdös-Rényi with $n = 100$ and $p_n = 6/100$



Threshold for isolated vertices I

Theorem

"Not having an isolated vertex" has threshold function $\frac{\log n}{n}$.

Proof: Let X_n be the number of isolated vertices in the Erdös-Rényi graph $G_n \sim \mathbb{G}_{n,p_n}$. Using $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$,

$$\mathbb{E}_{n,p_n}[X_n] = n(1-p_n)^{n-1} \leq e^{\log n - (n-1)p_n} \to 0,$$

when $p_n \gg \frac{\log n}{n}$. So the first moment method gives one direction: $\mathbb{P}_{n,p_n}[X_n > 0] \to 0$ when $p_n \gg \frac{\log n}{n}$.

Threshold for isolated vertices II

Proof (continued): Let A_j be the event that vertex j is isolated and $X_n = \sum_j \mathbf{1}_{A_j}$. By the computation above, using $1 - x \geq e^{-x-x^2}$ for $x \in [0, 1/2]$,

$$\mu_n = \mathbb{E}_{n,p_n}[X_n] = \sum_i \mathbb{P}_{n,p_n}[A_i] = n(1-p_n)^{n-1} \geq e^{\log n - np_n - np_n^2},$$

which goes to $+\infty$ when $p_n \ll \frac{\log n}{n}$.

Note that for all $i \neq j$

$$\mathbb{P}_{n,p_n}[A_i \cap A_j] = (1-p_n)^{2(n-2)+1},$$

so that

$$\gamma_n = \mathbb{E}_{n,p_n}[X_n^2] - \mathbb{E}_{n,p_n}[X_n] = \sum_{i \neq j} \mathbb{P}_{n,p_n}[A_i \cap A_j] = n(n-1)(1-p_n)^{2n-3}.$$

Threshold for isolated vertices III

Proof (continued): We have

$$\begin{aligned}\frac{\mathbb{E}_{n,p_n}[X_n^2]}{(\mathbb{E}_{n,p_n}[X_n])^2} &= \frac{\mu_n + \gamma_n}{\mu_n^2} \\ &\leq \frac{n(1-p_n)^{n-1} + n^2(1-p_n)^{2n-3}}{n^2(1-p_n)^{2n-2}} \\ &\leq \frac{1}{n(1-p_n)^{n-1}} + \frac{1}{1-p_n},\end{aligned}$$

which is $1 + o(1)$ when $p_n \ll \frac{\log n}{n}$. ■

Threshold for connectivity I

Theorem

Connectivity has threshold function $\frac{\log n}{n}$.

Proof: We start with the easy direction. If $p_n \ll \frac{\log n}{n}$, the previous result implies that the graph has isolated vertices, and therefore is disconnected, with probability going to 1 as $n \to +\infty$.

Now assume that $p_n \gg \frac{\log n}{n}$. Let $\mathcal{D}_n$ be the event that G_n is disconnected. To bound $\mathbb{P}_{n,p_n}[\mathcal{D}_n]$, for $k \in \{1, \ldots, n/2\}$ we let Y_k be the number of subsets of k vertices that are disconnected from all other vertices in the graph. Then, by the first moment method,

$$\mathbb{P}_{n,p_n}[\mathcal{D}_n] \leq \mathbb{P}_{n,p_n}\left[\sum_{k=1}^{n/2} Y_k > 0\right] \leq \sum_{k=1}^{n/2} \mathbb{E}_{n,p_n}[Y_k].$$

Threshold for connectivity II

Proof (continued): Using that $k \leq n/2$ and $\binom{n}{k} \leq n^k$,

$$\mathbb{E}_{n,p_n}[Y_k] = \binom{n}{k}(1-p_n)^{k(n-k)} \leq \left(n(1-p_n)^{n/2}\right)^k.$$

The expression in parentheses is $o(1)$ when $p_n \gg \frac{\log n}{n}$. Summing over k,

$$\mathbb{P}_{n,p_n}[\mathcal{D}_n] \leq \sum_{k=1}^{+\infty} \left(n(1-p_n)^{n/2}\right)^k = O(n(1-p_n)^{n/2}) = o(1),$$

where we used that the geometric series (started at $k = 1$) is dominated asymptotically by its first term. ∎

Go deeper

More details and examples on the first and second moment methods at:

`http://www.math.wisc.edu/~roch/mdp/`

For more on random graphs in general, see e.g. (available online):

- *Random Graphs and Complex Networks. Vol. I and II* by van der Hofstad
- *Random Graph Dynamics* by Durrett

Probability on Graphs: Techniques and Applications to Data Science

2 - Exponential tail bounds

Sébastien Roch
UW–Madison
Mathematics

July 26, 2018

Moment-generating function

Definition

The *moment-generating function* of X is the function

$$M_X(s) = \mathbb{E}\left[e^{sX}\right],$$

defined for all $s \in \mathbb{R}$ where it is finite, which includes $s = 0$.

Chernoff-Cramér bound

Assume X is a centered (i.e. mean 0) random variable such that $M_X(s) < +\infty$ for $s \in (-s_0, s_0)$ for some $s_0 > 0$. Exponentiating within Markov's inequality gives, for any $\beta > 0$ and $s > 0$,

$$\mathbb{P}[X \geq \beta] = \mathbb{P}[e^{sX} \geq e^{s\beta}] \leq \frac{M_X(s)}{e^{s\beta}} = \exp\left[-\left\{s\beta - \Psi_X(s)\right\}\right],$$

where $\Psi_X(s) = \log M_X(s)$. The best exponent is

$$\Psi_X^*(\beta) = \sup_{s \in \mathbb{R}_+} (s\beta - \Psi_X(s)).$$

Chernoff-Cramér for sums of independent variables

Let $S_n = \sum_{i \le n} X_i$, where the X_is are i.i.d. centered random variables. Then

$$\Psi_{S_n}(s) = \log \mathbb{E}[e^{s \sum_{i \le n} X_i}] = \log \prod_{i \le n} \mathbb{E}[e^{sX_i}] = n \Psi_{X_1}(s)$$

Theorem

Assume $M_{X_1}(s) < +\infty$ *on* $s \in (-s_0, s_0)$ *for some* $s_0 > 0$. *For any* $\beta > 0$,

$$\mathbb{P}[S_n \ge \beta] \le \exp\left(-n \Psi^*_{X_1}\left(\frac{\beta}{n} \right) \right).$$

Example: Binomial

Let Z_n be a binomial random variable with parameters n and p. Recall that Z_n is a sum of i.i.d. indicators $Y_1, \ldots, Y_n$ and, letting $X_i = Y_i - p$ and $S_n = Z_n - np$,

$$\Psi_{X_1}(s) = \log \mathbb{E}[e^{s(Y_1 - p)}] = \log\left(pe^s + (1-p)\right) - ps.$$

For $b \in (0, 1-p)$, letting $a = b + p$, direct calculation gives

$$\begin{aligned} \Psi^*_{X_1}(b) &= \sup_{s>0}(sb - (\log\left[pe^s + (1-p)\right] - ps)) \\ &= (1-a)\log\frac{1-a}{1-p} + a\log\frac{a}{p} =: D(a\|p), \end{aligned}$$

achieved at $s_b = \log\frac{(1-p)a}{p(1-a)}$. By the previous result, for $\beta > 0$,

$$\mathbb{P}[Z_n \geq np + \beta] \leq \exp\left(-n\, D\left(p + \beta/n \| p\right)\right).$$

Sub-Gaussian variables I

Let $X \sim N(0, \nu)$ where $\nu > 0$ and note that

$$\begin{aligned} M_X(s) &= \int_{-\infty}^{+\infty} e^{sx} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{x^2}{2\nu}} \, \mathrm{d}x = \int_{-\infty}^{+\infty} e^{\frac{s^2\nu}{2}} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{(x-s\nu)^2}{2\nu}} \, \mathrm{d}x \\ &= \exp\left(\frac{s^2\nu}{2}\right), \end{aligned}$$

so that straightforward calculus gives for $\beta > 0$

$$\Psi_X^*(\beta) = \sup_{s>0}(s\beta - s^2\nu/2) = \frac{\beta^2}{2\nu},$$

achieved at $s_\beta = \beta/\nu$. Plugging $\Psi_X^*(\beta)$ into Theorem 2 leads for $\beta > 0$ to the bound

$$\mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right).$$

Sub-Gaussian variables II

We say that a centered random variable X is *sub-Gaussian with variance factor* $\nu > 0$ if for all $s \in \mathbb{R}$

$$\Psi_X(s) \leq \frac{s^2 \nu}{2},$$

which is denoted by $X \in \mathcal{G}(\nu)$. By the Chernoff-Cramér bound

$$\mathbb{P}[X \leq -\beta] \vee \mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right),$$

where we used that $X \in \mathcal{G}(\nu)$ implies $-X \in \mathcal{G}(\nu)$.

Example: Back to the binomial

Theorem (Case $p = 1/2$)

Let $X_1, \dots, X_n$ be independent $\{-1, 1\}$-valued random variables with $\mathbb{P}[X_i = 1] = \mathbb{P}[X_i = -1] = 1/2$. Let $S_n = \sum_{i \le n} X_i$. Then, for any $\beta > 0$,

$$\mathbb{P}[S_n \ge \beta] \le e^{-\beta^2/2n}.$$

Proof: The moment-generating function of X_1 can be bounded as follows

$$M_{X_1}(s) = \frac{e^s + e^{-s}}{2} = \sum_{j \ge 0} \frac{s^{2j}}{(2j)!} \le \sum_{j \ge 0} \frac{(s^2/2)^j}{j!} = e^{s^2/2}. \tag{1}$$

So $\Psi_{S_n}(s) = n\Psi_{X_1}(s) \le s^2 n/2$ and $S_n \in \mathcal{G}(n)$. ■

Sub-Gaussian variables III

Theorem (General Hoeffding inequality)

Let $X_1, \dots, X_n$ be independent centered random variables with $X_i \in \mathcal{G}(\nu_i)$ for $0 < \nu_i < +\infty$ and let $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$. Let $S_n = \sum_{i \le n} \alpha_i X_i$. Then $S_n \in \mathcal{G}(\sum_{i=1}^n \alpha_i^2 \nu_i)$ and for all $\beta > 0$,

$$\mathbb{P}[S_n \ge \beta] \le \exp\left(-\frac{\beta^2}{2\sum_{i=1}^n \alpha_i^2 \nu_i}\right).$$

Proof: By independence,

$$\Psi_{S_n}(s) = \sum_{i \le n} \Psi_{\alpha_i X_i}(s) = \sum_{i \le n} \Psi_{X_i}(s\alpha_i) \le \sum_{i \le n} \frac{(s\alpha_i)^2 \nu_i}{2} = \frac{s^2 \sum_{i \le n} \alpha_i^2 \nu_i}{2}.$$

■

Example: Bounded variables I

For bounded random variables, the previous inequality reduces to a standard bound.

Theorem (Hoeffding's inequality)

Let $X_1, \dots, X_n$ be independent random variables where, for each i, X_i takes values in $[a_i, b_i]$ with $-\infty < a_i \leq b_i < +\infty$. Let $S_n = \sum_{i \leq n} X_i$. For all $\beta > 0$,

$$\mathbb{P}[S_n - \mathbb{E}S_n \geq \beta] \leq \exp\left(-\frac{2\beta^2}{\sum_{i \leq n}(b_i - a_i)^2}\right).$$

Illustration: Maximum degree of Erdös-Rényi

Let $G_n \sim \mathbb{G}_{n,p}$ be an Erdös-Rényi graph with n vertices and density $p_n = p \in (0,1)$. Let D_i be the degree of vertex i and let $D^* = \max_i D_i$. Note that D_i is $\mathrm{Bin}(n-1,p)$, i.e. a sum of independent $[0,1]$-variables, so by Hoeffding's inequality

$$\mathbb{P}_{n,p}[D_i - (n-1)p \geq \sqrt{(1+\varepsilon)n\log(n)/2}] \leq e^{-(1+\varepsilon)\log n}.$$

By a union bound

$$\begin{aligned}
&\mathbb{P}_{n,p}[D^* \geq (n-1)p + \sqrt{(1+\varepsilon)n\log(n)/2}] \\
&\quad \leq \sum_i \mathbb{P}_{n,p}[D_i - (n-1)p \geq \sqrt{(1+\varepsilon)n\log(n)/2}] \\
&\quad \leq n \times n^{-(1+\varepsilon)} \to 0.
\end{aligned}$$

Example: Bounded variables II

Proof: By the general Hoeffding inequality, it suffices to show that $X_i - \mathbb{E}X_i \in \mathcal{G}(\nu_i)$ with $\nu_i = \frac{1}{4}(b_i - a_i)^2$. We give a quick proof of a weaker version that uses a trick called *symmetrization*. Suppose the X_is are centered and satisfy $|X_i| \leq c_i$ for some $c_i > 0$. Let X_i' be an independent copy of X_i and let Z_i be an independent uniform in $\{-1, 1\}$. By Jensen's inequality

$$\mathbb{E}\left[e^{sX_i}\right] = \mathbb{E}\left[e^{s\mathbb{E}[X_i - X_i' \mid X_i]}\right] \leq \mathbb{E}\left[\mathbb{E}\left[e^{s(X_i - X_i')} \,\middle|\, X_i\right]\right] = \mathbb{E}\left[e^{s(X_i - X_i')}\right].$$

By the symmetry of $X_i - X_i'$, we then get

$$\begin{aligned}\mathbb{E}\left[e^{s(X_i - X_i')}\right] &= \mathbb{E}\left[e^{sZ_i(X_i - X_i')}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{sZ_i(X_i - X_i')} \,\middle|\, X_i, X_i'\right]\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[e^{(s(X_i - X_i'))^2/2} \,\middle|\, X_i, X_i'\right]\right] \leq \mathbb{E}\left[e^{(s(X_i - X_i'))^2/2}\right] \leq e^{-2c_i^2 s^2}.\end{aligned}$$

■

Many more concentration inequalities

- Bernstein's inequality
- Azuma's inequality
- Matrix inequalities

Epsilon-nets I

Exponential tail inequalities are useful, among other things, to study the deviations of suprema of random variables. When the supremum is over an *infinite* index set, one way to proceed is to apply a tail inequality to a sufficiently dense finite subset of the index set, and then extend the resulting bound by continuity. This is referred to as an ε*-net argument*.

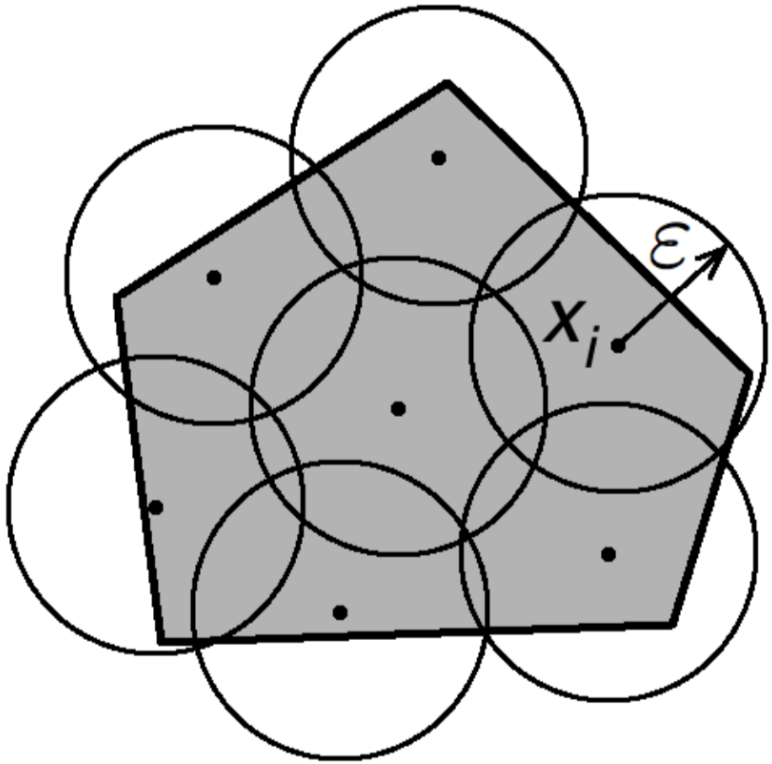
Epsilon-nets II

Definition (ε-net)

Let S be a subset of a metric space (M, ρ) and let $\varepsilon > 0$. A collection of points $N \subseteq S$ is called an *ε-net of S* if all pairs of points in N are at distance greater than ε and N is maximal by inclusion in S. In particular for all $z \in S$, $\inf_{y \in N} \rho(z, y) \leq \varepsilon$. The *covering number* of S, denoted by $\mathcal{N}(S, \rho, \varepsilon)$, is the smallest cardinality of an ε-net of S.

The definition of an ε-net immediately suggests an algorithm for constructing one. Start with $N = \emptyset$ and successively add a point to N at distance at least ε from all other previous points until that is not possible to do so anymore. (Provided S is compact, this procedure will terminate after a finite number of steps.)

Epsilon-nets by picture



(a) This covering of a pentagon K by seven ε-balls shows that $\mathcal{N}(K, \varepsilon) \leq 7$.

Illustration: Spectral norm of random matrix I

For a $m \times n$ matrix $A \in \mathbb{R}^{m \times n}$, recall that the spectral norm is defined as

$$\|A\| := \sup_{\mathbf{x} \in \mathbb{R}^n \setminus \{0\}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x} \in \mathbb{S}^{n-1}} \|A\mathbf{x}\|_2 = \sup_{\substack{\mathbf{x} \in \mathbb{S}^{n-1} \\ \mathbf{y} \in \mathbb{S}^{m-1}}} \langle A\mathbf{x}, \mathbf{y} \rangle,$$

where $\mathbb{S}^{n-1}$ is the sphere of radius 1 around the origin in $\mathbb{R}^n$.

(To see the rightmost equality above, note that Cauchy-Schwarz implies $\langle A\mathbf{x}, \mathbf{y} \rangle \leq \|A\mathbf{x}\|_2 \|\mathbf{y}\|_2$ and that one can take $\mathbf{y} = A\mathbf{x}/\|A\mathbf{x}\|_2$ for any $\mathbf{x}$ such that $A\mathbf{x} \neq 0$ in the rightmost expression.)

Illustration: Spectral norm of random matrix II

Theorem

Let $A \in \mathbb{R}^{m \times n}$ be a random matrix whose entries are centered, independent and sub-Gaussian with variance factor ν. Then there exist a constant $0 < C < +\infty$ such that, for all $t > 0$,

$$\|A\| \leq C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t),$$

with probability at least $1 - e^{-t^2}$.

Without independence of the entries, the spectral norm can be much larger. Say A is all-$(+1)$ or all-(-1) with equal probability. Taking the vector $\mathbf{x} = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$ shows that $\|A\| \geq n$ with probability 1.

Illustration: Spectral norm of random matrix III

Proof: We seek to bound

$$\|A\| = \sup_{\substack{\mathbf{x}\in\mathbb{S}^{n-1}\\ \mathbf{y}\in\mathbb{S}^{m-1}}} \langle A\mathbf{x}, \mathbf{y}\rangle = \sup_{\substack{\mathbf{x}\in\mathbb{S}^{n-1}\\ \mathbf{y}\in\mathbb{S}^{m-1}}} \sum_{i,j} x_i y_j A_{ij},$$

where we note that the last quantity is a linear combination of independent variables. Fix $\varepsilon = 1/4$. We proceed in two steps:

1. We first apply the general Hoeffding inequality to control the deviations of the supremum *restricted to* ε*-nets* N *and* M *of* $\mathbb{S}^{n-1}$ *and* $\mathbb{S}^{m-1}$.
2. We then extend the bound to the full supremum by continuity.

Back to ε-nets: Sphere

Let $\mathbb{S}^{k-1}$ be the sphere of radius 1 centered around the origin in $\mathbb{R}^k$ with the Euclidean metric. Let $0 < \varepsilon < 1$. We claim that

$$\mathcal{N}(S, \rho, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^k.$$

Let N be any ε-net of S. The balls of radius $\varepsilon/2$ around points in N, $\{\mathbb{B}^k(x_i, \varepsilon/2) : x_i \in N\}$, satisfy two properties:

1. Pairwise disjoint: if $z \in \mathbb{B}^k(x_i, \varepsilon/2) \cap \mathbb{B}^k(x_j, \varepsilon/2)$, then $\|x_i - x_j\|_2 \leq \|x_i - z\|_2 + \|x_j - z\|_2 \leq \varepsilon$, a contradiction.
2. Contained in $\mathbb{B}^k(0, 3/2)$: if $z \in \mathbb{B}^k(x_i, \varepsilon/2)$, then $\|z\|_2 \leq \|z - x_i\|_2 + \|x_i\| \leq \varepsilon/2 + 1 \leq 3/2$.

The volume of a ball of radius is $\varepsilon/2$ is $\frac{\pi^{k/2}(\varepsilon/2)^k}{\Gamma(k/2+1)}$ and that of a ball of radius $3/2$ is $\frac{\pi^{k/2}(3/2)^k}{\Gamma(k/2+1)}$. Divide one by the other.

Illustration: Spectral norm of random matrix IV

Lemma

Let N and M be as above. For C large enough, for all $t > 0$,

$$\mathbb{P}\left[\max_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2} C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right] \leq e^{-t^2}.$$

Proof: By the general Hoeffding inequality, $\langle A\mathbf{x}, \mathbf{y}\rangle$ is sub-Gaussian with variance factor

$$\sum_{i,j}(x_i y_j)^2 \nu = \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \nu = \nu,$$

for all $\mathbf{x} \in N$ and $\mathbf{y} \in M$. In particular, for all $\beta > 0$,

$$\mathbb{P}\left[\langle A\mathbf{x}, \mathbf{y}\rangle \geq \beta\right] \leq \exp\left(-\frac{\beta^2}{2\nu}\right).$$

Illustration: Spectral norm of random matrix V

Proof of lemma (continued): Hence, by a union bound over N and M,

$$\begin{aligned}
&\mathbb{P}\left[\max_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x},\mathbf{y}\rangle \geq \frac{1}{2}C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right] \\
&\quad\leq \sum_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \mathbb{P}\left[\langle A\mathbf{x},\mathbf{y}\rangle \geq \frac{1}{2}C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right] \\
&\quad\leq |N||M| \exp\left(-\frac{1}{2\nu}\left\{\frac{1}{2}C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right\}^2\right) \\
&\quad\leq 12^{n+m} \exp\left(-\frac{C^2}{8}\left\{m+n+t^2)\right\}\right) \\
&\quad\leq e^{-t^2},
\end{aligned}$$

for $C^2/8 = \log 12 \geq 1$, where in the third inequality we ignored all cross-products since they are non-negative. ∎

Illustration: Spectral norm of random matrix VI

Lemma

For any ε-nets N and M of $\mathbb{S}^{n-1}$ and $\mathbb{S}^{m-1}$ respectively, the following inequalities hold

$$\sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \leq \|A\| \leq \frac{1}{1-2\varepsilon} \sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle.$$

Proof: The first inequality is immediate. For the second inequality, we will use the following observation

$$\langle A\mathbf{x}, \mathbf{y}\rangle - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle = \langle A\mathbf{x}, \mathbf{y}-\mathbf{y}_0\rangle + \langle A(\mathbf{x}-\mathbf{x}_0), \mathbf{y}_0\rangle.$$

Fix $x \in \mathbb{S}^{n-1}$ and $y \in \mathbb{S}^{m-1}$ such that $\langle A\mathbf{x}, \mathbf{y}\rangle = \|A\|$, and let $\mathbf{x}_0 \in N$ and $\mathbf{y}_0 \in M$ such that

$$\|\mathbf{x}-\mathbf{x}_0\|_2 \leq \varepsilon \qquad \text{and} \qquad \|\mathbf{y}-\mathbf{y}_0\|_2 \leq \varepsilon.$$

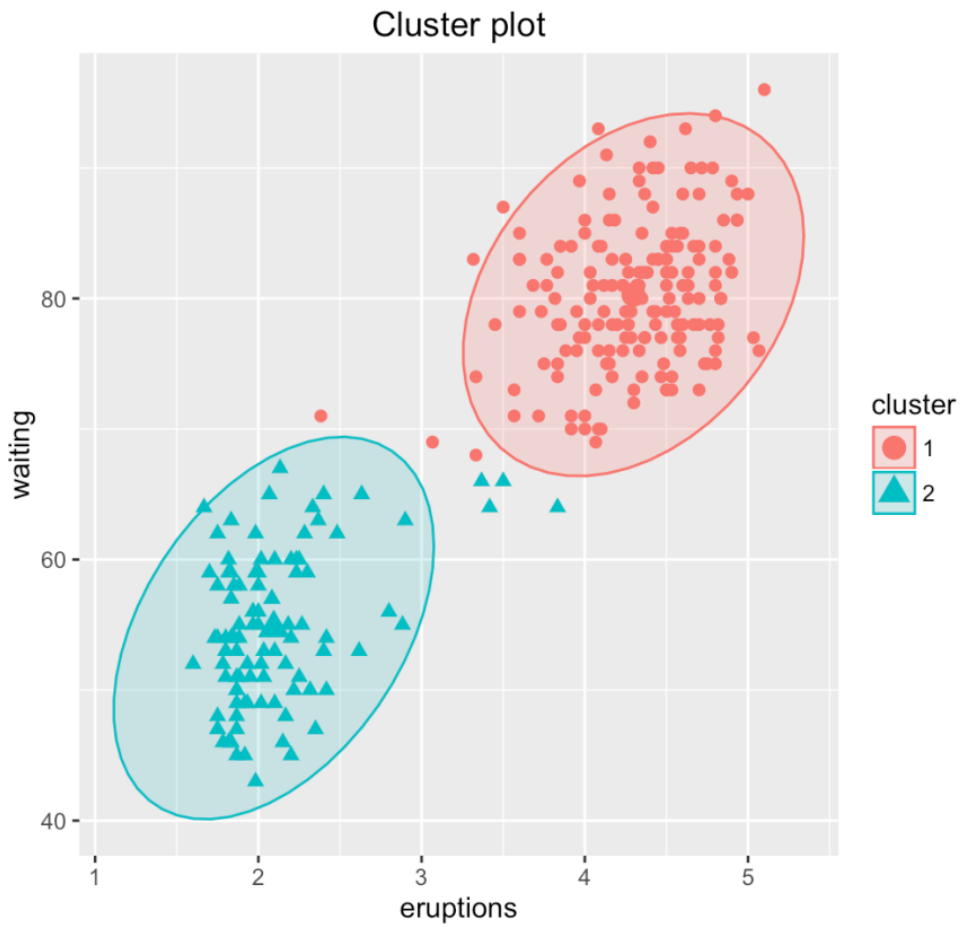
Illustration: Spectral norm of random matrix VII

Proof of lemma (continued): Then the inequality above, Cauchy-Schwarz and the definition of the spectral norm imply

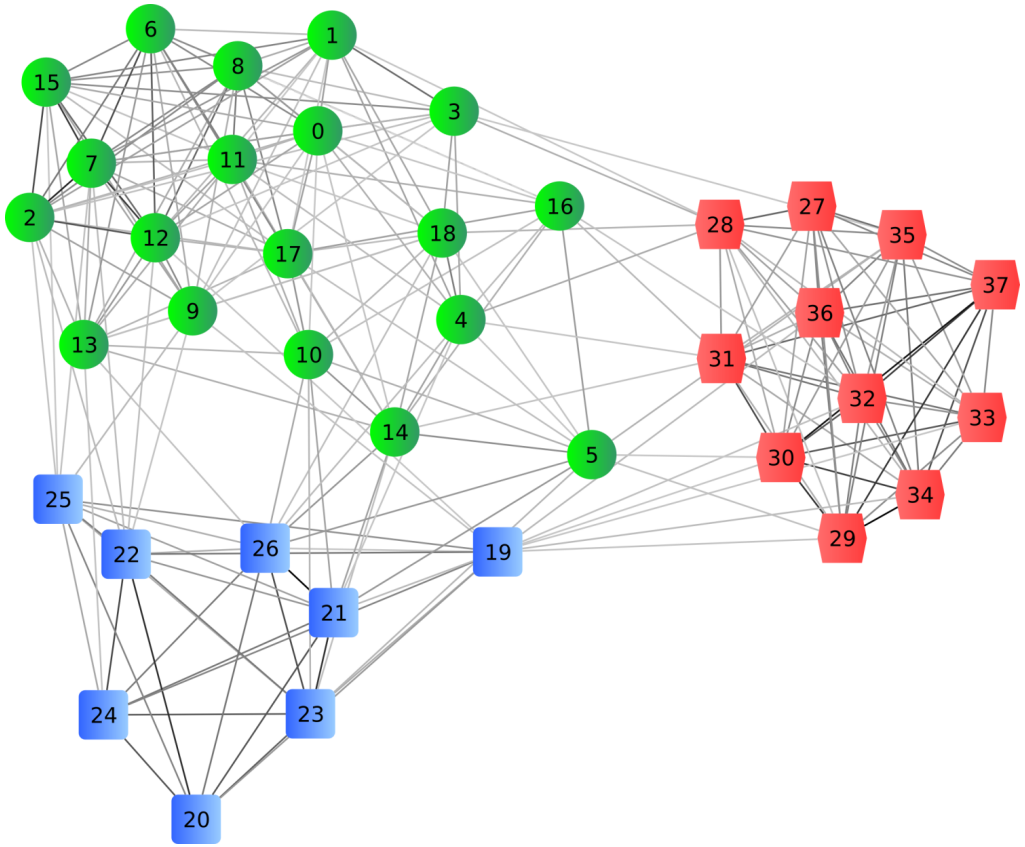
$$\|A\| - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle \leq \|A\|\|\mathbf{x}\|_2\|\mathbf{y} - \mathbf{y}_0\|_2 + \|A\|\|\mathbf{x} - \mathbf{x}_0\|_2\|\mathbf{y}_0\|_2 \leq 2\varepsilon\|A\|.$$

Rearranging gives the claim. ■

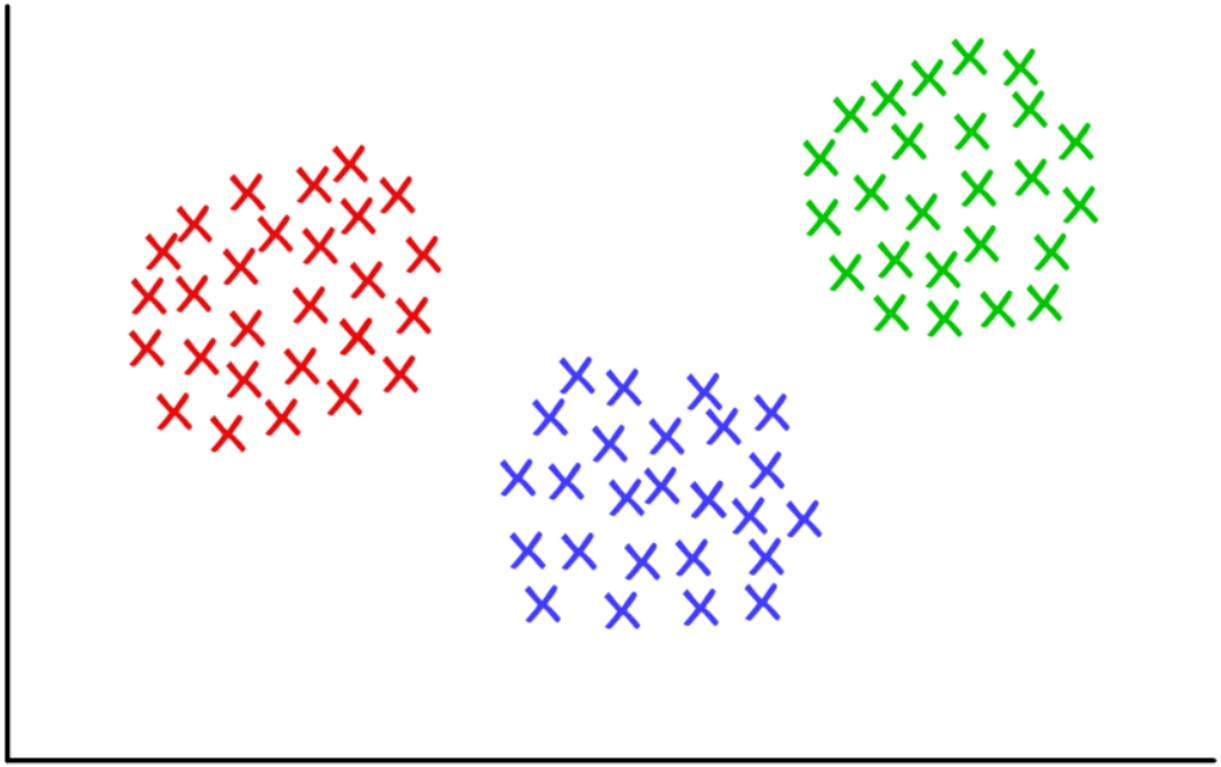
Clustering in Euclidean space



Clustering in graphs



Reducing the second problem to the first one



Stochastic blockmodel with two balanced blocks

Definition

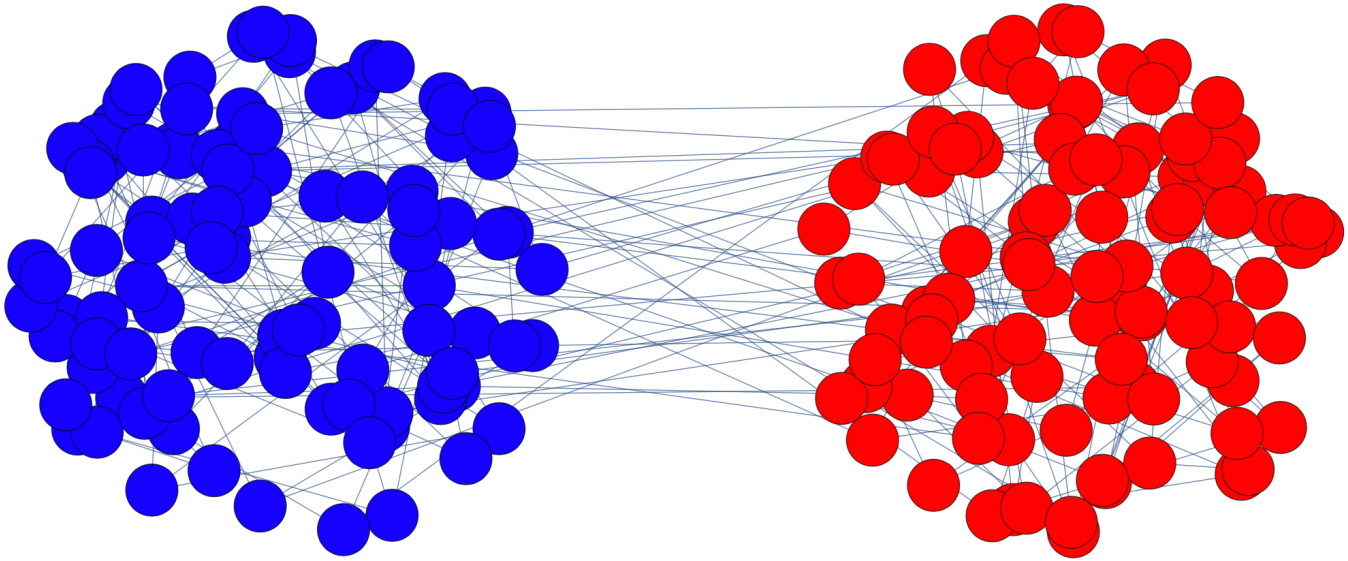
Let $V = [n]$ with n even, let $V_1 = \{1, \dots, n/2\}$ and $V_2 = \{n/2 + 1, \dots, n\}$, and let $0 < q < p < 1$. We draw a graph $G = (V, E)$ at random as follows. For each pair $x \neq y$ in V, the edge $\{x, y\}$ is in E with probability:

- p if $x, y \in V_1$, or $x, y \in V_2$;
- q if $x \in V_1$ and $y \in V_2$, or $x \in V_2$ and $y \in V_1$;

independently of all other edges. We write $G \sim \mathrm{SBM}_{n,p,q}$ and we denote the corresponding measure by $\mathbb{P}_{n,p,q}$.

Community detection problem: Given G (without the node labels), output V_1, V_2 (possibly approximately).

Stochastic blockmodel by picture



Expected adjacency matrix

Let $G \sim \mathrm{SBM}_{n,p,q}$ and let A be the adjacency matrix of G.

Theorem

Let $D = \mathbb{E}_{n,p,q}[A]$. *Then*

$$D = n\frac{p+q}{2}\mathbf{u}_1\mathbf{u}_1^T + n\frac{p-q}{2}\mathbf{u}_2\mathbf{u}_2^T - p\,I,$$

where $\mathbf{u}_1 = \frac{1}{\sqrt{n}}(1,\ldots,1)^T$ *and* $\mathbf{u}_2 = \frac{1}{\sqrt{n}}(1,\ldots,1,-1,\ldots,-1)^T$.

Proof: Note that D is a block matrix with diagonal blocks all-p and off-diagonal blocks all-q, all of size $n/2 \times n/2$, with the exception of the diagonal which is all-0. ■

Idea: Compute the second eigenvector of A and cluster by sign.

Spectral clustering: a positive result

Theorem

Let $G \sim \mathrm{SBM}_{n,p,q}$ and let A be the adjacency matrix of G. Let $\mu = \min\left\{q, \frac{p-q}{2}\right\} > 0$. Clustering according to the sign of the second eigenvector of A identifies the two communities of G with probability at least $1 - e^{-n}$, except for C/μ^2 misclassified nodes for some constant $C > 0$.

Matrix perturbation

Theorem (A version of Davis-Kahan)

Let S and T be symmetric $n \times n$ matrices. Let $\lambda_i(S)$ be the i-th largest eigenvalue of S with corresponding unit eigenvector $\mathbf{v}_i(S)$ (and similarly for T). If

$$\delta := \min_{j \neq i} |\lambda_i(S) - \lambda_j(S)| > 0,$$

then there is $\theta \in \{+1, -1\}$ such that

$$\|\mathbf{v}_i(S) - \theta\, \mathbf{v}_i(T)\|_2 \leq \frac{4\|S - T\|}{\delta}.$$

Bounding the spectral norm

Lemma

Let $G \sim \mathrm{SBM}_{n,p,q}$*, let* A *be the adjacency matrix of* G *and let* $D = \mathbb{E}_{n,p,q}[A]$*. Then, there is a constant* $C > 0$ *such that*

$$\|A - D\| \leq C\sqrt{n},$$

with probability at least $1 - e^{-n}$*.*

Proof: The entries of R are centered, independent and sub-Gaussian with variance factor $1/4$. ∎

Spectral clustering: proof I

Proof of spectral clustering theorem: The eigenvalues of D are

$$n\frac{p+q}{2} - p, \qquad n\frac{p-q}{2} - p, \qquad -p,$$

so $\lambda_2(D) = n\frac{p-q}{2} - p$ and

$$\delta = \min_{j\neq 2} |\lambda_2(D) - \lambda_j(D)| = \min\left\{n\frac{p-q}{2}, nq\right\} =: n\mu > 0.$$

By Davis-Kahan and the previous lemma, with probability at least $1 - e^{-n}$, there is $\theta \in \{+1, -1\}$ such that

$$\|\mathbf{v}_2(D) - \theta\,\mathbf{v}_2(A)\|_2 \leq \frac{4C\sqrt{n}}{n\mu} \leq \frac{C'}{\sqrt{n}\,\mu}.$$

Spectral clustering: proof II

Proof of spectral clustering theorem (continued): Put differently,

$$\sum_i \left|\sqrt{n}\,(\mathbf{v}_2(D))_i - \sqrt{n}\,\theta\,(\mathbf{v}_2(A))_i\right|^2 \leq \frac{(C')^2}{\mu^2}.$$

If the signs of $(\mathbf{v}_2(D))_i$ and $\theta\,(\mathbf{v}_2(A))_i$ disagree, then the i-th term in the sum above is ≥ 1. So there can be at most $(C')^2/\mu^2$ of those. That establishes the desired bound on the number of misclassified nodes. ■

Go deeper

More details and examples on tail bounds at:

`http://www.math.wisc.edu/~roch/mdp/`

For more on concentration in general, see e.g. (available online):

- *High-dimensional probability: An introduction with applications in data science* by Vershynin
- *Probability in High Dimension* by van Handel