Contents

Part I

Network Models

Chapter 1

Branching Processes

1.1 General setting

X random variable in $\mathbb{N}_0$.
$c = \mathbb{E}[X]$, $p_k = \Pr(X = k)$.
D_X: offspring distribution.

Build a random tree T_X inductively

- Start with root.
- Add $X_1 \sim D_X$ children.
- Given generation $t-1$ with $Z_{t-1} = k$ individuals, add $X_{1,i}$ children to the ith individual in generation $t-1$, where $X_{t,1}, \ldots, X_{t,k} \sim D_X$ i.i.d.

Expectation of Z_n Assume $Z_{n-1} = k$. Then: $Z_n = \sum_{i=1}^{k} X_{n,k}$ and $\mathbb{E}[Z_n] = k\,\mathbb{E}[X] = kc$.

$$\begin{aligned} \mathbb{E}[Z_n \big| Z_{n-1}] &= cZ_{n-1} \\ \mathbb{E}[Z_n] &= c\,\mathbb{E}[Z_{n-1}] = c^2\,\mathbb{E}[Z_{n-2}] = \cdots = c^n \end{aligned} \tag{1.1}$$

Note that $|T_X| = \sum_{n=0}^{\infty} Z_n$, proving that:

$$\mathbb{E}[|T_X|] = \begin{cases} \frac{1}{1-c} & c < 1 \quad \text{subcritical} \\ \infty & c = 1 \quad \text{critical} \\ \infty & c > 1 \quad \text{supercritical} \end{cases} \tag{1.2}$$

Survival probability $\theta = \Pr(|T_X| = \infty)$

Theorem 1.1. Let $G(\lambda) = \sum_k p_k \lambda^k$ be the PGF of X. Then $\eta = 1 - \theta$ is the smallest solution of $\lambda = G(\lambda)$ in $[0, 1]$.

Consequences Note that $G', G'' > 0, G(0) = p_0, G(1) = 1$, and $G'(\lambda) = \sum_k p_k k \lambda^{k-1}$ gives:

$$G'(1) = \mathbb{E}[X] = c$$

Proof that $\eta = G(\eta)$. Condition on $X_1 = k$, i.e. the root having k children.

We have k subtrees $T_1, \ldots, T_k$. $|T| < \infty \Leftrightarrow |T_1| < \infty, \ldots |T_k| < \infty$.

Conditioned on $X_1 = k$, $T_1, \ldots, T_k$ are independent and have the same distribution as the original tree T; in particular, $\Pr(|T_i| < \infty) = \eta$.

Thus, we have:

$$\begin{aligned}
\eta &= \Pr(|T| < \infty) \\
&= \sum_k \Pr(X = k) \Pr(|T_1| < \infty, \ldots |T_k| < \infty) \\
&= \sum_k \Pr(X = k) \prod_{i=1}^{k} \Pr(|T_i| < \infty) \qquad (1.3) \\
&= \sum_k \Pr(X = k) \eta^k \\
&= \mathbb{E}_X[\eta^X] = G_X(\eta)
\end{aligned}$$

□

1.2 Random walk representation

Idea: explore T_X via BFS.

At time t, D_t: set of discovered vertices, Q_t: queue, y_t: length of queue.

$\boxed{t = 0}$: $D_0 = Q_0 = \{v_1\}$

$\boxed{t = 0 \to t = 1}$: $Q_1 = \{v_2, \ldots, v_{1+X_1}\}, D_1 = \{v_1\} \cup Q_1, y_t = y_0 - 1 + X_1 = X_1$

- dequeue $v_1 \in Q_0$
- fing the children of v_1
- v_1 has X_1 children $v_2, \ldots, v_{1+X_1}$, where $X_1 \sim D_X$
- add $v_2, \ldots, v_{1+X_1}$ to the queue

$\boxed{t - 1 \to t}$ if $Q_{t-1} \neq \emptyset$: $y_t = y_{t-1} - 1 + X_t, D_t = v_1, \ldots, v_t \cup Q_t$

- dequeue $v_t \in Q_{t-1}$
- find its children (there are $X_t \sim D_X$)
- add them to the queue

Analysis

1. If $y_t > 0 \quad \forall t$, then $|T_X| = \infty$

2. Let $k = \min\{t : y_t = 0\}$, i.e. last leaf of T_X. Then $D_k = \{v_1, \ldots, v_k\} \cup Q_k = \{v_1, \ldots v_k\}$ and $|T_X| = k$.
3. As long as $y_{t-1} > 0$, $y_t = 1 + X_1 + \cdots + X_t - t \geq 0$, $X_1 + \cdots + X_t = t - 1 + y_t$; in particular:

$$X_1 + \cdots + X_{t-1} = t - 2 + y_{t-1} \geq t - 1$$

Thus:

$$\boxed{|T| = k \Leftrightarrow k = \min\{t : y_t = 0\} \Leftrightarrow X_1 \geq 1, X_1 + X_2 \geq 2, \ldots, X_1 + \cdots + X_{k-1} \geq k - 1, \sum_{i=1}^{k} X_i = k - 1}$$

Random walk drift Define drift as $\mathbb{E}[y_t - y_{t-1}] = \mathbb{E}[X_t] - 1 = c - 1$. Note that we have $\mathbb{E}[y_t] = 1 + (c-1)t$ that drifts up or down, depending on $c > 1$ or $c < 1$.

Key: very unlikely to die very late. May still die because of fluctuations.

We expect $1 + X_1 + \ldots X_t - t = 1 + t(\mathbb{E}[X] - 1) \pm \sqrt{t}$, thus fluctuations should lead to it hitting zero at some time t:

$$|T_X| = \min\{t : y_t = 0\} = \min\{t : 1 + X_1 + \cdots + X_t - t = 0\} < \infty$$

where the precise hitting time is random.

The standard deviation is $\sqrt{\mathrm{var}(y_t)} = \sqrt{\mathrm{var}(\sum_{i=1}^{t}(X_i - 1))} = \sqrt{t\,\mathrm{var}(X_t)} \approx \sqrt{t}$.

Theorem 1.2. If $\Pr(X = 1) = 1$, then $\theta = 1$. If $\Pr(X = 1) < 1$, then $\theta > 0$ is equivalent to $c > 1$.

Implicit equation $\eta = G(\eta)$.

Lemma 1.1. If extinction probability $\eta = 1 - \theta = \Pr(|T_X| < \infty)$, then η satisfies $\eta = G_X(\eta)$.

Theorem 1.3. For general $X \in \mathbb{N}_0$, $\eta = 1 - \theta$ is the smallest solution of $G_X(\lambda) = \lambda$.

Corollary 1.1. If $c > 1$, $\Pr(k \leq |T_X| < \infty) \leq \frac{e^{-Dk}}{1 - e^{-D}}$.

Proof.

$$\begin{aligned}
\Pr(k \leq |T_X| < \infty) &\leq \sum_{l=k}^{\infty} \Pr(|T_X| = k) \leq \sum_{l=k}^{\infty} e^{-Dk} \\
&\leq \frac{1}{1 - e^{-D}} - \frac{e^{-D \cdot 0}(1 - e^{-D})^{k-1}}{1 - e^{-D}} \\
&\leq \frac{e^{-Dk}}{1 - e^{-D}}
\end{aligned}$$

□

Corollary 1.2. If $c > 1$, $\theta = \Pr(|T_X| = \infty) > 0$.

Proof by contradiction. Assume $\theta = 0$. Then, we have:

$$\begin{aligned}\mathbb{E}[|T_X|] &= \sum_{k\geq 0} k \Pr(|T_X| = k) \\ &\leq \sum_{k\geq 0} k e^{-Dk} \quad e^{-D} \in (0,1) \\ &= \frac{e^{-D}}{(1-e^{-D})^2} < \infty\end{aligned}$$

But $\mathbb{E}[|T_X|] = \mathbb{E}[\sum_{t=0}^{\infty} Z_t] = \sum_{t\geq 0} \mathbb{E}[Z_t] = \sum_{t\geq 0} c^t = \infty$. Contradiction. □

Lemma 1.2 (Homework 3)**.** If $c > 1$ and $\delta \in (0,1)$, then $\Pr(c^{(1-\delta)n} \leq Z_t \leq c^{(1+\delta)n} \big| |T_X| = \infty) \to 1$.

Proof. RHS: Use Markov's inequality. **Key:** exponential growth above criticality.

$$\Pr(Z_n > c^{(1+\delta)n}) \leq \frac{\mathbb{E}[Z_n]}{c^{(1+\delta)n}} = c^{-\delta n}$$

i.e.

$$\Pr(Z_n > c^{(1+\delta)n} \big| |T_X| = \infty) = \frac{\Pr(Z_n > c^{(1+\delta)n} \cap |T_X| = \infty)}{\theta} \leq \frac{1}{\theta}\frac{1}{c^{\delta n}} \to 0$$

LHS: Choose $c' = c^{1-\delta/2} < c$.

$$\Pr(Z_{n+1} \leq c't \big| Z_n = t) = \Pr(\sum_{i=1}^{t} X_i \leq c't) \leq e^{-tD}$$

where $D = I(c') = \sup_{t\leq 0}(c't - \log \mathbb{E}[e^{tX}])$.
Use the law of total probability:

$$\begin{aligned}\Pr(Z_{n+1} \leq c'k \cap Z_n \geq k) &= \sum_{t=k}^{\infty} \Pr(Z_{n+1} \leq c'k \cap Z_n = t) \\ &= \sum_{t=k}^{\infty} \Pr(Z_{n+1} \leq c'k \big| Z_n = t) \Pr(Z_n = t) \\ &= \sum_{t=k}^{\infty} \Pr(Z_{n+1} \leq c't \big| Z_n = t) \Pr(Z_n = t) \quad \text{relax constraints} \\ &\leq \sum_{t=k}^{\infty} e^{-tD} \Pr(Z_n = t) \\ &\leq e^{-kD} \Pr(Z_n \geq k) \leq e^{-kD}\end{aligned}$$

$$\Pr(Z_n < c^{(1-\delta)n} \big| |T_X| = \infty) \leq \Pr(Z_{n_0+t} < c^{(1-\delta/2)t} k_0 \big| |T_X| = \infty) \leq \epsilon$$

□

Lemma 1.3. Let $X \sim \text{Poiss}(c)$ or $\text{Bin}(n, \frac{c}{n})$.

a) If $c < 1$, $\Pr(|T_X| > k) \leq e^{-\frac{(1-c)^2}{2}k}$

b) If $c > 1$, $\Pr(|T_X| = k) \le e^{-\frac{(c-1)^2}{2c}k}$

Concentration bound for Poisson random variables Given $X_i \overset{\text{i.i.d.}}{\sim} \text{Poiss}(c)$.

$$
\begin{aligned}
\Pr(\sum_{i=1}^{k} X_i \ge \lambda k) &\le e^{-\frac{(\lambda-c)^2}{2\lambda}k} \quad \text{if } \lambda > c \\
\Pr(\sum_{i=1}^{k} X_i \le \lambda k) &\le e^{-\frac{(c-\lambda)^2}{2c}k} \quad \text{if } \lambda < c
\end{aligned}
\tag{1.4}
$$

Proof of first part.

$$
\begin{aligned}
\Pr(\sum_{i=1}^{k} X_i \ge \lambda k) &\le \frac{\mathbb{E}[e^{t\sum_{i=1}^{k} X_i}]}{e^{\lambda t k}} = \mathbb{E}[e^{tX}]^k e^{-\lambda t k} = e^{-k(\lambda t - c(e^t - 1))} \quad \text{maximize over } t \text{ to get } I(\lambda) \\
&\le e^{-kI(\lambda)}
\end{aligned}
$$

Define rate function $I(\lambda) = \lambda \ln(\frac{\lambda}{c}) - \lambda + c$:

$$
\begin{aligned}
I'(x) &= \ln(\frac{x}{c}) \\
I''(x) &= \frac{1}{x}
\end{aligned}
$$

Use Taylor expansion with remainder:

$$
I(\lambda) = I(c) + I'(c)(\lambda - c) + \frac{I''(x)}{2!}(\lambda - c)^2 = \frac{(\lambda - c)^2}{2x} \ge \frac{(\lambda - c)^2}{2\lambda}
$$

where $x \in [c, \lambda]$. □

Multiplicative Chernov bound

$$
\begin{aligned}
\Pr(S_k \ge (1+\epsilon)\,\mathbb{E}[S_k]) &\le e^{-\frac{\epsilon^2}{2(1+\epsilon)}\,\mathbb{E}[S_k]} \\
\Pr(S_k \le (1-\epsilon)\,\mathbb{E}[S_k]) &\le e^{-\frac{\epsilon^2}{2}\,\mathbb{E}[S_k]}
\end{aligned}
\tag{1.5}
$$

Lemma 1.4. If S_k is a sum of k i.i.d. random variables which are either exponential random variables with expectation 1, Poisson random variables with arbitrary expectation, or arbitrary random variables taking values in $[0, 1]$. Then S_k obeys the multiplicative Chernov bound.

Proof.

(i) Let $T_1, \ldots, T_k \sim \text{Exp}(1)$ i.i.d. and $S_k = \sum_{i=1}^{k} T_i$.

Note that for $t \in (0, 1)$ we have $\mathbb{E}[e^{tT_i}] = \frac{1}{1-t} = e^{-\log(1-t)}$.

Then, we have:

$$\begin{aligned}\Pr(S_k \geq (1+\epsilon)k) &= \Pr(e^{tS_k} \geq e^{(1+\epsilon)tk}) \\ &\leq e^{-(1+\epsilon)tk}e^{-k\log(1-t)} = e^{-k((1+\epsilon)t+\log(1-t))}\end{aligned}$$

Similarly, we have:

$$\begin{aligned}\Pr(S_k \leq (1-\epsilon)k) &= \Pr(e^{-tS_k} \geq e^{-(1-\epsilon)tk}) \quad \mathbb{E}[e^{-tT_i}] = \frac{1}{1+t} \quad \forall t \in (0,\infty) \\ &\leq e^{(1-\epsilon)tk}e^{-k\log(1+t)} \quad 1-\epsilon = \frac{1}{1+t^*}, t^* = \frac{\epsilon}{1-\epsilon} \\ &\leq e^{(1-\epsilon)\epsilon k + k\log(1-\epsilon)} \quad \log(1-\epsilon) \leq -\epsilon \quad \forall \epsilon < 1 \\ &= e\end{aligned}$$

$\square$

Concentration bound for general random variables Given X a random variable with $\mathbb{E}[X] = c \in \mathbb{R}$, $X_i \overset{\text{i.i.d.}}{\sim} D_X$.

$$\Pr(\sum_{i=1}^{k} X_i \geq \lambda k) \leq e^{-kI_X(\lambda)} \quad \text{if } \lambda > c$$

$$\Pr(\sum_{i=1}^{k} X_i \leq \lambda k) \leq e^{-kI_X(\lambda)} \quad \text{if } \lambda < c$$

where $I_X(\lambda)$ is the *rate function*

$$I_X(\lambda) = \begin{cases} \sup_{t\geq 0}(t\lambda - \log\mathbb{E}[e^{tX}]) & \text{if } \lambda \geq c \\ \sup_{t\leq 0}(t\lambda - \log\mathbb{E}[e^{tX}]) & \text{if } \lambda \leq c \end{cases} \tag{1.6}$$

Lemma 1.5. If $\mathbb{E}[X] = c > 1$, there exists constant $D > 0$ s.t. $\Pr(|T_X| = k) \leq e^{-Dk}$.

Key: A supercritical BP either survives or dies out fast.

Proof.

$$\begin{aligned}\Pr(|T_X| = k) &\leq \Pr(X_1 + \cdots + X_k < k) \\ &= \Pr(e^{-t\sum_{i=1}^{k} X_i} \geq e^{-tk}) \quad \text{Chernoff bound} \\ &\leq \mathbb{E}[e^{-tX}]^k e^{tk} \\ &= e^{-k(-t-\log\mathbb{E}[e^{-tX}])} \\ &\leq e^{-kI(1)} \quad \lambda = 1\end{aligned}$$

where $I(1) = \sup_{t\leq 0}\{1 \cdot t - \log\mathbb{E}[e^{tX}]\} = \sup_{\zeta\geq 0}\{-\zeta - \log\mathbb{E}[e^{-\zeta X}]\} = \sup_{\zeta\geq 0}\phi(\zeta)$.

$$\phi(0) = 0$$

$$\phi'(\zeta) = -1 + \frac{\mathbb{E}[Xe^{-\zeta X}]}{\mathbb{E}[e^{-\zeta X}]} \to \mathbb{E}[X] - 1 = c - 1 > 0 \quad \text{as } \zeta \to 0$$

Choose $D = I(1) = \sup_{\zeta \geq 0} \phi(\zeta) > 0$. $\square$

Rate function See Cramer's theorem lecture.

1.3 Duality

Theorem 1.4. Consider a branching process with offspring distribution $\Pr(X = k) = p_k$ and extinction probability $\eta \in (0, 1)$. Then the original branching process conditioned on extinction is a branching process with probability distribution $p_k^* = \eta^{k-1} p_k$.

Key:

- Survival requires at least one infinite lineage; extinction requires all branches finite
- Conditioning on extinction biases reproduction toward smaller families.

General duality theorem Let X have PGF $G(s) = \sum_{k\ geq0} p_k s^k$ and extinction probability η solving $\eta = G(\eta)$.

Theorem 1.5 (Duality theorem)**.** Conditioning on eventual extinction yields another Galton-Watson process with PGF

$$G^*(s) = \frac{G(\eta s)}{\eta}$$

which is subcritical and dies out with probability 1.

Key: a supercritical process, given extinction, looks subcritical.

Why the factor η^{k-1} appears If the root has k children:

- all k subtrees must die: η^k
- conditioning on extinction: η^{-1}

With reproduction weight p_k, the conditioned offspring probability is proportional to $p_k \eta^{k-1}$, giving:

$$G^*(s) = \mathbb{E}[s^X] = \sum_{k \geq 0} p_k \eta^{k-1} s^k = \frac{1}{\eta} \sum_{k \geq 0} p_k (\eta s)^k = \frac{G(\eta s)}{\eta}$$

Proof by induction. Let $E = \{\text{eventual extinction}\}$ and $\eta = \Pr(E)$.

Base: For the root, $\eta = \sum_k p_k \eta^k = G(\eta)$.

Inductive hypothesis: At some generation, conditioned offspring weights are $p_k \eta^{k-1}$.

Inductive step: A parent with k children has all subtrees die with probability η^k; conditioning on E divides by η, yielding weights $p_k \eta^{k-1}$ and PGF $G^*(s) = G(\eta s)/\eta$. Thus, the form persists one generation deeper.

Key: every time we enforce finiteness on another level, we multiply by η; the single division by η comes from conditioning on extinction overall. $\square$

Problem 1.1. Show that if $X \sim \text{Poiss}(c)$ witn $c > 1$, the offspring distribution p_k^* is again Poisson with parameter $\tilde{c} = c\eta$, and show that

$$ce^{-c} = \tilde{c}e^{-\tilde{c}}$$

Prove that the function $c \mapsto f(c) = ce^{-c}$ has a unique max at $c = 1$ and is strictly monotone increasing up to max, strictly monotone decreasing after that, and $f(c) \to 0$ as $c \to 0$ and $c \to \infty$.

Conclude that for any $c > 1$, $ce^{-c} = \tilde{c}e^{-\tilde{c}}$ has a unique solution $\tilde{c} < 1$.

Solution: Use duality theorem.

Chapter 2

Thresholds in Random Graphs

2.1 Erdos-Renyi random graphs

$G(n,p)$: each of $\Pi = \binom{n}{2}$ possible edges is chosen with probability p.

Definition 2.1 (Graph property)**.** A graph property P is invariant under relabelling.

- sample space Ω: set of all graphs on $[n]$
- property $P \subseteq \Omega$: all graphs with $d = 1$; containing a triangle; not containing a 4-cycle

2.2 Thresholds

Notation If $a_n, b_n \geq 0$.

$$a_n \ll b_n \Leftrightarrow a_n/b_n = o(1) \Leftrightarrow a_b/b_n \to 0$$
$$a_n \gg b_n \Leftrightarrow a_n/b_n \to \infty$$

Definition 2.2. $\rho(n)$ is a threshold for P iff

$$\Pr(P) \to 1 \quad \text{if } p \gg \rho(n)$$
$$\Pr(P) \to 0 \quad \text{if } p \ll \rho(n)$$

or if this holds for $P^C =$ not P.

Proof strategy Let X random variable in $\mathbb{N}_0$: number of edges, k-trees, k-cycles.

1. First moment method: show $\mathbb{E}[X] \to 0$ if $p \ll \rho(n)$ using $\Pr(X > 0) \leq E[X]$
2. Second moment method: show $\mathbb{E}[X^2] \leq \mathbb{E}[X]^2(1 + \epsilon_n)$ where $\epsilon \to 0$ if $p \gg \rho(n)$

Lemma 2.1. If $X \geq 0$ then $\Pr(X > 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}$.

Proof. Use Cauchy inequality: $\mathbb{E}[X] = \mathbb{E}[X] \cdot \mathbb{E}[\mathbb{1}_{X>0}] \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[\mathbb{1}_{X>0}]}$. $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2] \Pr(X > 0)$. $\square$

Edges Let X be the number of edges in $G(n,p)$, E: set of edges, $\Pi = \binom{n}{2}$: number of possible edges.
$X = \sum_e \mathbb{1}_{e\in E}$.

$$\mathbb{E}[X] = \sum_e \mathbb{E}[\mathbb{1}_{e\in E}] = \sum_e \Pr(e \in E) = \Pi p \to 0 \quad \text{if } p \ll n^{-2}$$

$$\mathbb{E}[X^2] = \sum_{e,e'\in E} \mathbb{E}[\mathbb{1}_{e\in E}\mathbb{1}_{e\in E}] = \sum_{e=e'} + \sum_{e\neq e'} = \Pi p + \Pi(\Pi-1)p^2 \leq \Pi p + \Pi^2 p^2 = \mathbb{E}[X]^2(1 + \frac{1}{\Pi p})$$

Choose $\epsilon_n = \frac{1}{\Pi p}$: $\Pr(X > 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} \geq \frac{1}{1+\epsilon_n} \to 1$ if $p \gg n^{-2}$.

k-trees $X = \sum_{T:|V(T)|=k} \mathbb{1}_{T\subseteq G(n,p)}$.

$$\mathbb{E}[X] = \sum_T \mathbb{E}[\prod_{e\in T} \mathbb{1}_{e\in E}] = \sum_T p^{k-1} = \sum_{\substack{S\subseteq[n]\\|S|=k}} \sum_{T:V(T)=S} p^{k-1} = \sum_{\substack{S\subseteq[n]\\|S|=k}} k^{k-2} p^{k-1} = k^{k-2}p^{k-1}\binom{n}{k}$$
$$\leq \frac{k^{k-2}}{k!} p^{k-1} n^k \to 0 \quad \text{if } p \ll n^{-\frac{k}{k-1}}$$

Cayley's formula: number of trees on n vertices is n^{n-2}.

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{T,T'} \mathbb{E}[\mathbb{1}_{T\subseteq G(n,p)}\mathbb{1}_{T'\subseteq G(n,p)}] \\
&= \sum_{T,T'} p^{|E(T)\cup E(T')|} \\
&= \sum_{T,T'} p^{2k-2-|E(T)\cap E(T')|} \\
&\leq \sum_{T,T'} p^{2k-2} + \sum_{\substack{T,T'\\|V(T)\cap V(T')|=\zeta}} p^{2k-2-(\zeta-1)} \\
&\leq \mathbb{E}[X]^2 + \sum_{\zeta=1}^{k} \sum_{\substack{|S|=|S'|=k\\|S\cap S'|=\zeta}} (k^{k-2}p^{k-1})^2 p^{-(\zeta-1)} \\
&\leq \mathbb{E}[X]^2 + \left(k^{k-2}p^{k-1}\binom{n}{k}\right)^2 C(k) \sum_{\zeta=2}^{k} \binom{k}{\zeta} n^{-\zeta} p^{-(\zeta-1)} \\
&\leq \mathbb{E}[X]^2 \left(1 + C(k) 2^k \max_{\zeta=2,\dots,k} n^{-\zeta} p^{-(\zeta-1)}\right) = \mathbb{E}[X]^2(1+\epsilon_n)
\end{aligned}$$

Then, we have $\Pr(X > 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} \geq \frac{1}{1+\epsilon_n}$. If $p \gg n^{-\frac{k}{k-1}}$, then $p \gg n^{-\frac{\zeta}{\zeta-1}}$ and $p^{\zeta-1}n^{\zeta} \gg 1$; $\epsilon_n \to 0$.

Key: if $|E(T)\cap E(T')| = l \geq 1$, then $|V(T)\cap V(T')| \geq l+1$.

- S: $\binom{n}{k}$
- $S \cap S'$: $\binom{k}{\zeta}$
- $S \setminus S'$: $\binom{n-k}{k-\zeta} \leq n^{k-\zeta}$

Note:

$$\frac{n^k}{\binom{n}{k}} \leq C(k) = k!k^k; \quad \sum_{\zeta=2}^{k} \binom{k}{\zeta} \leq 2^k$$

k-cycles Let X_k: a number of k-cycles, N_k: a number of k-cycles on k vertices $\leq k!$.

$$\begin{aligned}\mathbb{E}[X_k] &= \sum_C \Pr(C \subseteq G(n,p)) \\ &= \sum_{S:|S|=k} \sum_{C \text{ on } S} p^k \\ &= \binom{n}{k} N_k p^k \leq \frac{n^k}{k!} k! p^k = (np)^k \to \quad \text{if } p \ll \frac{1}{n}\end{aligned}$$

Lemma 2.2. Assume that $p \ll \frac{1}{n}$. Then w.h.p. the connected components of $G(n,p)$ are trees of size $o(\log n)$.

Proof. Assume $k \geq 2$, $npe \leq 1/4$. Let Y_k: number of cycles of size k, Y: number of cycles, X_k: number of trees of size k.

$$\mathbb{E}[X_k] \leq \frac{1}{epk^2}(npe)^k$$

$$\mathbb{E}[X_{\geq k}] \leq \sum_{l \geq k} \frac{1}{epl^2}(npe)^l \leq \frac{1}{epk^2} \frac{(npe)^k}{1-npe} \leq \frac{2}{epk^2}(npe)^k \leq n^2pe(npe)^{k-2} \approx pn^2(npe)^{k-2}$$

$\square$

Isolated vertices Let X be the number of isolated vertices. If $i \in [n]$ isolated, none of the edges between i and $[n]\imath$ is in $G(n,p)$:

$$\Pr(i \text{ is isolated}) = (1-p)^{n-1}$$

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i \in [n]} \Pr(i \text{ is isolated}) \\ &= n(1-p)^{n-1} \leq ne^{-p(n-1)} = nee^{-np} \quad p \geq (1+\epsilon)\frac{\log n}{n} \\ &\leq ne\frac{1}{n^{1+\epsilon}} = \frac{e}{n^\epsilon} \to 0\end{aligned}$$

Definition 2.3. $\rho(n)$ is a sharp threshold for P iff

$$\Pr(P) \to \begin{cases} 0 & \text{if } p \leq (1-\epsilon)\rho(n) \\ 1 & \text{if } p \geq (1+\epsilon)\rho(n) \end{cases}$$

Problem 2.1. Show that:

- $\frac{\log n}{n}$ is a sharp threshold after which there are no isolated vertices
- $\frac{\log n}{n}$ is a threshold for $G(n,p)$ being connected

Chapter 3

Giant Component in Random Graphs

For $p \ll \frac{1}{n}$, there are no cycles, i.e. just isolated trees.

3.1 Component structure

Theorem 3.1. If $np \to \infty$, then with high probability $\exists$ a giant component $C_{\max}$ with $|C_{\max}| \geq n(1 - e^{-\frac{np}{4}})$ and all other components are trees.

Proof. With high probability:

a) all components C with $|C| \leq n/2$ are trees

b) if X_k: number of trees of size $k \leq \frac{n}{2}$, then $\mathbb{E}[X_k] \leq \frac{1}{pk^2}((pne)e^{-\frac{np}{2}})^k$

c) if Z_k: number of components of size k, $Z = \sum_k kZ_k$: number of total vertices in small components, then $\mathbb{E}[Z] = O(ne^{-\frac{np}{2}})$ and

$$\Pr(Z \geq ne^{-\frac{np}{4}}) = O(e^{-\frac{np}{4}}) \to 0$$

d) number of vertices in component $|C| > n/2$ is $n - Z$

but there can only be one component with number of vertices greater than $n/2$. $\square$

Solution:

a) Let X_k: number of sets S s.t. $|S| = k$, the induced graph is a component, S contains at least k edges.

$$\begin{aligned}
\Pr(S \text{ contributes}) &= \Pr(\exists T \text{ on } S \text{ and one edge}) \\
&\leq \sum_{T \text{ on } S} \sum_{\text{extra edge}} p^{k-1} p (1-p)^{k(n-k)} \\
&\leq k^{k-2} k^2 p^k e^{-pk(n-k)} \qquad (k \leq n/2) \\
&\leq (kp)^k e^{-knp/2}
\end{aligned}$$

$$\mathbb{E}\left[\sum_k X_k\right] = \sum_{k=3}^{n/2} \binom{n}{k} (kp)^k e^{-knp/2}$$
$$\leq \sum_{k\geq 3} \frac{k^k}{k!} (npe^{-np/2})^k \quad k! \geq \sqrt{2\pi k}\left(\frac{k}{e}\right)^k \geq \left(\frac{k}{e}\right)^k$$
$$\leq \sum_{k\geq 3} (npee^{-np/2})^k \to 0 \quad \text{if } np \to \infty$$

i.e. there all connected components of size less than or equal to $n/2$ are trees if $np \to \infty$.

b) Let X_k: number of isolated trees of size k.

- $k-1$ edges: p^{k-1}
- $\binom{k}{2} - (k-1)$ internal non-edges: $(1-p)^{\binom{k}{2}-(k-1)}$
- $k(n-k)$ non-edges isolating T from the other $n-k$ verticess: $(1-p)^{k(n-k)}$

$$\mathbb{E}[X_k] = \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{\binom{k}{2}-(k-1)+k(n-k)} \quad \binom{k}{2} - (k-1) + k(n-k) \geq nk/2$$
$$\leq \frac{n^k}{k!} k^{k-2} p^{k-1} e^{-pnk/2}$$
$$\leq \frac{1}{pk^2} ((pne)e^{-np/2})^k$$

c)

$$\mathbb{E}[Z] = \sum_{k\leq n/2} k\,\mathbb{E}[X_k] \leq \frac{1}{p}\sum_k (pnee^{-np/2})^k \leq \frac{1}{p}\frac{pnee^{-np/2}}{1-pnee^{-np/2}} = O(\frac{1}{p}pne^{-np/2}) = O(ne^{-np/2})$$

3.2 Giant in $G(n,p)$

Key idea $\Pr(\text{vertex } v \in C_{\max}) = \Pr(\text{BFS from vertex } v \text{ survives}) = \theta$

$$\mathbb{E}[|C_{\max}|] = \mathbb{E}\left[\sum_v \mathbb{1}_{v \in C_{\max}}\right] = \sum_v \Pr(\text{vertex } v \in C_{\max}) = n\theta$$

Notation **Key:** For BP with offspring distribution $\sim \mathrm{Poi}(c)$, survival $\theta = 1 - \eta = 1 - e^{c(\eta-1)} = 1 - e^{-c\theta}$

$G(n,p)$: n nodes, each of $\binom{n}{2}$ possible edges, chosen i.i.d. with probability $p = \frac{c}{n}$ for constant c.
$C(x)$: connected component that contains vertex x, $C_{\max}$: largest connected component.
$\theta = \Pr(|T_{\mathrm{Poiss}(c)}| = \infty)$.

Theorem 3.2 (Phase transition at $c = 1$)**.**

a) If $c < 1$, w.h.p. $|C_{\max}| = O(\log n)$.

b) If $c > 1$, $\frac{|C_{\max}|}{n} \to \theta$ and w.h.p. $|C(x)| = O(\log n)$ for all other components.

Proof strategy. **Key:** $|C(x)| \le k \Leftrightarrow \exists l \le k : X_1 + \cdots + X_k \le l - 1$

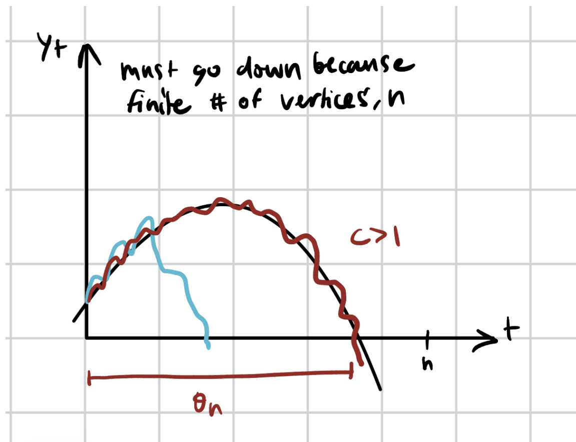
- Explore $C(x)$ via BFS
- Relate the probability that $x \in C_{\max}$ to the probability $\theta(c)$ that a $\mathrm{Poiss}(c)$ branching process survives
- Similar to the argument which shows that a supercritical BP either survives or dies out fast, show that for $c > 1$, either $x \in C_{\max}$ or $|C(x)| = O(\log n)$

$\square$

Exploration of $C(x)$

- start with $v_1 = x,\ Q_0 = \{x\}$
- v_1 has $n-1$ possible children, so $X_1 \sim \mathrm{Bin}(n-1, p)$
  actual children in BFS tree $Q_1 = \{v_2, \ldots, v_{X_1+1}\},\ y_1 = X_1$
- v_2 has $n - 1 - X_1$ possible children, so $X_2 \sim \mathrm{Bin}(n-1-X_1, p)$
  actual children in BFS tree $y_2 = y_1 - 1 + X_2$ if $y_1 > 0$
- $y_t = \begin{cases} y_{t-1} - 1 + X_t & \text{if } y_{t-1} > 0 \\ 0 & \text{if } y_{t-1} = 0 \end{cases}$
  $|C(x)| = k \Leftrightarrow k = \min\{t : y_t = 0\}$
  $\Leftrightarrow X_1 \ge 1, X_1 + X_2 \ge 2, \ldots, X_1 + \cdots + X_{k-1} \ge k-1$ and $X_1 + \cdots + X_k = k - 1$
  but now $X_t \sim \mathrm{Bin}(n - 1 - X_1 - \cdots - X_t, p) = \mathrm{Bin}(N_{t-1}, p)$
- drift

$$\mathbb{E}[y_t \big| y_{t-1}] - y_{t-1} = \mathbb{E}[X_t] - 1 = \frac{c}{n} N_{t-1} - 1 < c - 1$$



Lemma 3.1 (Comparison lemma)**.** $T_{n,p}$: branching process with $\text{Bin}(n,p)$ offspring.

$$\Pr(|T_{n-k,p}| > k) \le \Pr(|C(x)| > k) \le \Pr(|T_{n,p}| > k)$$

Corollary 3.1. If $c < 1$, then w.h.p.

$$\max_{x \in [n]} |C(x)| \le \frac{2}{(1-c)^2} \log n$$

Proof.

$$\begin{aligned}
\Pr(|C(x)| > k) &= \Pr(|T_{\text{Poi}(c)}| > k) \\
&\le \Pr(|T_{\text{Bin}(n,c/n)}| > k) \quad \Leftrightarrow y_t = 1 + (X_1 + \cdots + X_t) - t > 0, t \le k \Leftrightarrow \sum_{i=1}^{k} X_i \ge k, X_i \sim \text{Bin}(n, c/n) \\
&= \Pr\left(\sum_{i=1}^{k} X_i \ge k\right) \\
&= \Pr(e^{t \sum_{i=1}^{k} X_i} \ge e^{tk}) \quad \mathbb{E}[e^{tX_i}] = \left(1 + \frac{c}{n}(e^t - 1)\right)^n \le e^{c(e^t-1)}, 1 + x \le e^x \; \forall x \in \mathbb{R} \\
&\le \frac{e^{tc(e^t-1)}}{e^{tk}} \\
&= e^{-k(t - c(e^t-1))} \\
&= e^{-k(-\log c + c - 1)} = e^{-kI(c)} \\
&\le e^{-\frac{(1-c)^2}{2}k} \quad k = \frac{3}{(1-c)^2} \log n \\
&\le n^{-3/2}
\end{aligned}$$

yielding $\Pr\left(\max_x |C(x)| > k\right) = nn^{-3/2} = n^{-1/2} = \frac{1}{\sqrt{n}} \to 0$. □

Consider $-\log c = -\log(1-(1-c)) = -\log(1-u)$. Then, we have $I(c) = -\log(1-u) - u$ and Taylor expansion around 0 yields:

$$-\log(1-u) - u = \frac{u^2}{2} + \frac{u^3}{3} + \cdots \ge \frac{u^2}{2}$$

where $u = 1 - c > 0$ for $c < 1$.

Corollary 3.2. If $c > 1$ and $k_n = o(n)$, $k_n \to \infty$

$$\Pr(|C(x)| > k_n) \to \Pr(|T_{\mathrm{Poi}(c)}| > k_n)$$

Proof.

$$\begin{aligned}\Pr(|C(x)| > k) &\le \Pr(|T_{n,p}| > k)\\ &= \Pr(k < |T_{n,p}| < \infty) + \Pr(|T_{n,p}| = \infty)\\ &\to O(e^{-kD}) + \theta(c)\end{aligned}$$

$$\begin{aligned}\Pr(|C(x)| > k) &\ge \Pr(|T_{n-k,p}| > k)\\ &\ge \Pr(|T_{n-k,p}| = \infty)\\ &\to \theta(c)\end{aligned}$$

We note that all components with $|C(x)| \ge \frac{3c}{(1-c)^2}\log n$ must be part of $C_{\max}$.

□

Corollary 3.3. Let k be a constant, $c > 1$, $V_t = \{x : |C(x)| > k\log n\}$. Then

$$\mathbb{E}\left[\frac{|V_t|}{n}\right] \to \theta$$

3.3 $G(n,p)$ has small diameter

Choose $x, y \in C_{\max}$ uniformly at random. Let $D = \mathrm{dist}_{G(n,p)}(x,y)$.
Claim: $D \approx \log_c n$

Theorem 3.3. For $c > 1$, $\frac{D}{\log_c n} \to 1$ in probability.

Proof. Need to show $\forall \epsilon > 0$

a) $\Pr(D \le (1-\epsilon)\log_c n) \to 0$ in probability

b) $\Pr(D \ge (1+\epsilon)\log_c n) \to 0$ in probability

□

Proof of a). Let $l_0 = (1-\epsilon)\log_c n$.
Claim: $\forall x \neq y \in [n]$

$$\Pr_{G(n,p)}(d(x,y) \le l_0) \le \frac{1}{n}\frac{c^{l_0}}{1-1/c} = \frac{n^{-\epsilon}}{1-1/c}$$

Proof: Let X_l: the number of paths $w : x \to y$ of length l

$$\begin{aligned}\Pr(\mathrm{dist}(x,y) = l) &\le \mathbb{E}[X_l]\\ &= p^l \cdot \text{number of paths}\\ &\le p^l n^{l-1}\\ &= \frac{1}{n}c^l\end{aligned}$$

Claim: a) holds.

$$\begin{aligned}
\Pr(D \le l_0) &= \Pr\left(D \le l_0 \cap |C_{\max}| \ge \frac{\theta}{2}n\right) + \cancel{\Pr\left(D \le l_0 \cap |C_{\max}| < \frac{\theta}{2}n\right)} \\
&= \mathbb{E}\left[\mathbb{1}_{|C_{\max}| \ge \frac{n\theta}{2}} \cdot \mathbb{1}_{D \le l_0}\right] \\
&= \mathbb{E}_{G(n,p)}\left[\mathbb{1}_{|C_{\max}| \ge \frac{n\theta}{2}} \cdot \frac{1}{|C_{\max}|^2} \sum_{x,y \in C_{\max}} \mathbb{1}_{\mathrm{dist}(x,y) \le l_0}\right] \qquad \text{conditioned on } G \sim G(n,P) \\
&\le \frac{4}{n^2\theta^2} \mathbb{E}_{G(n,p)}\left[\sum_{x,y\in[n]} \mathbb{1}_{D \le l_0}\right] \qquad \frac{1}{|C_{\max}|^2} \le \frac{4}{(n\theta)^2} \\
&= \frac{4}{n^2\theta^2} \sum_{x,y\in[n]} \Pr_{G(n,p)}(\mathrm{dist}(x,y) \le l_0) \\
&= \frac{4}{\theta^2 n} + \frac{4}{\theta^2 n^2} \sum_{x \ne y} \Pr_{G(n,p)}(\mathrm{dist}(x,y) \le l_0) \\
&= O\left(\frac{1}{n}\right) + O\left(\frac{1}{n^\epsilon}\right) \to 0
\end{aligned}$$

□

Proof of b).

Relationship between y_t and Z_k Let $t(k)$: time when last vertex of generation $k-1$ leaves queue. Then, we have $y_{t(k)} = Z_k$: everyone in queue.

$$t(1) = 1 : Z_1 = X_1 = y_1$$
$$\boxed{t(k+1) = t(k) + Z_k = t(k) + y_{t(k)}}$$

Heuristic $y_t \approxeq (c-1)t$

Then, we have $t(k+1) \approxeq t(k) + (c-1)t(k) \approxeq ct(k)$.

Set $t_0 = \frac{1+\epsilon}{2}\log_c n$.

Thus, we have:

$$y_{t_0} \approx c^{t_0} \approxeq n^{\frac{1+\epsilon}{2}}$$
$$\Pr(x \leftrightarrow y) \ge 1 - (1-p)^{n^{1+\epsilon}} \to 1$$

□

Chapter 4

Power laws 1: Configuration Model

Molloy Reed model $UG(\underset{\sim}{d})$

Given $\underset{\sim}{d} = (d_1, \ldots, d_n)$, draw G uniformly at random from all G with $d_i(G) = d_i \quad \forall i = 1, \ldots, n$

Configuration model $CM(\underset{\sim}{d})$

Given $\underset{\sim}{d} = (d_1, \ldots, d_n)$ with $l_n = \sum_i d_i$ even

- form set H of l_n half edges
- create $\mathcal{M}$: uniformly random matching into $l_n/2$ pairs
- output multipgraph $G = G(\mathcal{M})$

Remark 4.1. If we condition on $G \sim CM(\underset{\sim}{d})$ to be simple, we have $G \sim UF(\underset{\sim}{d})$

Assumptions:

- (D) there exists limiting r.v. $D \in \mathbb{N}$ s.t. $\frac{1}{n} \sum_i \mathbb{1}_{d_i = d} \to \Pr(D = d)$
- $(M1)$ $\frac{1}{n} \sum_i d_i \to \mathbb{E}[D] < \infty$
- $(M2)$ $\frac{1}{n} \sum_i d_i^2 \to \mathbb{E}[D^2] \leq \infty$

s.t. D: degree distribution of a random vertex.

$$\sum_i \frac{d_i}{l_n}(d_i - 1) = \frac{\sum_i d_i(d_i - 1)}{l_n} \qquad \frac{d_i}{l_n} : \text{reach vertex } i, d_i - 1 : \text{forward edges from vertex } i$$
$$= \frac{\sum_i d_i(d_i - 1)}{\sum_i d_i} \to \nu = \frac{\mathbb{E}[D(D-1)]}{\mathbb{E}[D]}$$

s.t. ν: weighted average number of forward edges from a vertex we reach via a random edge.

Theorem 4.1.

$$\Pr(G \text{ is simple}) \to e^{-\frac{\nu}{2} - \frac{\nu^2}{4}}$$

i.e. no self-loops.

Proof heuristic.

$$\begin{aligned}
\Pr(\Pi_{ii} = 0) &= \left(1 - \frac{d_i - 1}{l_n - 1}\right)\left(1 - \frac{d_i - 2}{l_n - 3}\right) \dots \left(1 - \frac{1}{l_n - 2(d_i - 2) - 1}\right) \\
&\approxeq \prod_{k=1}^{d_i - 1} \left(1 - \frac{d_i - k}{l_n}\right) \\
&\approxeq \exp\left(-\sum_{k=1}^{d_i - 1} \frac{d_i - k}{l_n}\right) \\
&= \exp\left(-\frac{d_i(d_i - 1)}{2l_n}\right) \qquad (d_i - 1)\dots 1 = \binom{d_i}{2}
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
\Pr(\exists \Pi_{ii} \neq 0) &\approxeq \prod_i \Pr(\Pi_{ii} \neq 0) \\
&\approxeq \exp\left(-\sum_i \frac{d_i(d_i - 1)}{2l_n}\right) \\
&\to e^{-\frac{\nu}{2}}
\end{aligned}$$

$\square$

4.1 Properties

Theorem 4.2. If $\min_i d_i \geq 3$, then w.h.p. $G \sim CM(\underset{\sim}{d})$ is connected.

Theorem 4.3. If $\Pr(D = 2) < 1$, then $G \sim CM(\underset{\sim}{d})$ has a unique giant with $\lim \frac{|C_{\max}|}{n} > 0$ i.e. $\nu > 1$.

Key: $\nu > 1$ means BP does not die out, and we keep branching.

Theorem 4.4. If $1 < \nu < \infty$ and x, y are chosen uniformly at random in $C_{\max}$, then

$$\frac{\text{dist}(x, y)}{\log_\nu n} \to 1 \quad \text{in probability}$$

Example 4.1. Power law graphs $\Pr(D = k) \approx k^{-\alpha}$

- $2 < \alpha < 3$: $\nu = \infty$, unique giant, $\text{dist}(x, y) \approxeq \log \log n$
- $\alpha > 3, \nu \leq 1$: no giant
- $\alpha > 3, \nu > 1$: unique giant, $\text{dist}(x, y) \approxeq \log_\nu n$

Why does $\mathbb{E}[D^2]$ enter? What is $\lim \frac{|C_{max}|}{n}$?

How many children has j? Number of children is equal to forward degree $d_j - 1$.
As the BFS continues, new connections are again formed uniformly among half-edge, i.e. proportional to the degree d_v of the chosen vertex v.

4.2 Two-step branching process

Root has k children with probability

$$p_k = \frac{1}{n}\sum_i \mathbb{1}_{d_i=k} \to \Pr(D=k)$$

Later vertices have degree $k+1$, and thus k children with probability

$$p_k^* = \frac{(k+1)\sum_i \mathbb{1}_{d_i=k+1}}{l_n} = \frac{(k+1)p_{k+1}}{\bar{d}}$$

where $\bar{d} = \frac{1}{n}\sum_i d_i = \frac{l_n}{n}$.

Definition 4.1. Let D be a r.v. on $\mathbb{N}_0$ with $0 < \mathbb{E}[D] < \infty$. Then T_{D,D^*} is the BP where the root has a random number D of children. and every other vertex has D^* children, where

$$\Pr(D^*=k) = \frac{(k+1)\Pr(D=k+1)}{\mathbb{E}[D]}$$

Note:

$$\begin{aligned}
\mathbb{E}[D^*] &= \sum_k k\Pr(D^*=k) \\
&= \frac{1}{\mathbb{E}[D]}\sum_k k(k+1)\Pr(D=k+1) \\
&= \frac{1}{\mathbb{E}[D]}\sum_k k(k-1)\Pr(D=k) \\
&= \frac{\mathbb{E}[D(D-1)]}{\mathbb{E}[D]}
\end{aligned}$$

Extinction probability

$$\begin{aligned}
\eta &= \Pr(|T_{D,D^*}| < \infty) \\
&= \sum_k \Pr(D=k)\Pr(|T_{D^*}| < \infty)^k \\
&= \boxed{G_D(\Pr(|T_{D^*}| < \infty))}
\end{aligned}$$

where

$$\begin{aligned}
\eta^* &= \Pr(|T_{D^*}| < \infty) \\
&= \sum \Pr(D^*=k)\Pr(|T_{D^*}| < \infty)^k \\
&= G_{D^*}(\eta^*)
\end{aligned}$$

We can show that the extinction probability is the smallest solution.

Proposition 4.1. The survival probability θ of T_{D,D^*} is equal to $1-\eta = 1-G_D(\eta^*)$ where η^* is the smallest solution of $\eta^* = G_{D^*}(\eta^*)$. Furthermore, if $\Pr(D=2) < 1$, then $\eta > 0 \iff \mathbb{E}[D^*] > 1$.

Theorem 4.5. Consider $CM(\underset{\sim}{d})$ s.t. $\Pr(D=2) < 1$. Then w.h.p. $\frac{|C_{\max}|}{n} \to \theta$ in probability, and all other components are of size $o(n)$. In particular, $\exists$ giant $\iff \mathbb{E}[D^*] > 1$.

Proof strategy.

1) Show
$$\Pr(|C(k)| > k) \to \Pr(|T_{D,D^*}| > k)$$

2) Show $\exists K_n \to \infty$ sufficiently slowly s.t.
$$\frac{|V_t|}{n} \to \theta \quad \text{in probability}$$
where $V_t = \{x : |C(x)| > k_n\}$.

3) Show that V_t is "mostly" connected, i.e. w.h.p we have
$$\max_{x \in V_t} |C(x)| \geq \theta n - o(n)$$

□

4.3 Self-loops and multi-loops

$$M_{ij} \leq \frac{1}{2} A_{ij}(A_{ij} - 1)$$
$$\mathbb{E}[A_{ij}(A_{ij} - 1)] =$$

Chapter 5

Power laws 2: Preferential Attachment

Key idea: there is a limited number of vertices that have significantly larger degree than rest.

- start with some finite graph G_{t_0} connected
- time $t \to t+1$: Given G_t with degrees $d_i(t)$:
  - add new vertex v
  - attach m edges i.i.d. $vu_1, vu_2, \ldots, vu_m$
- from v to $u_i \in G_t$, where $u_i = u$ with probability $P_u(t) = \frac{d_u(t)}{2|E(G_t)|}$

> Barabási-Albert '99 (or B-A) model has three versions:
>
> a) independent model: i.i.d. for $i = 1, \ldots, m$
>
> b) conditional model: condition on $u_i \neq u_j \quad \forall i \neq j$
>
> c) sequential model: update degrees in each step

Note: at each time step $t \to t+1$, we add one new vertex and m new edges

$$
\begin{aligned}
|V_t| &= |V_{t_0}| + t - t_0 \\
|E_t| &= |E_{t_0}| + (t - t_0)m
\end{aligned}
$$

Expected degrees

- time 1: $d_1(1) = d_2(1) = m$
- time $t \to t+1$: $d_i(t+1) - d_i(t) \big| d_i(t) \sim \text{Bin}\left(m, \frac{d_i(t)}{2mt}\right)$

$$
\mathbb{E}[d_i(t+1)|G_t] = d_i(t) + m\frac{d_i(t)}{2mt} = d_i(t)\left(1 + \frac{1}{2t}\right)
$$

Vertex i is born at time t_i with $d_i = m$: $t_1 = 1,\ t_i = i - 1 \quad i \geq 2$

$$\begin{aligned}
\mathbb{E}[d_i(t)] &= d_i(t_i)\prod_{\zeta=t_i}^{t-1}\left(1+\frac{1}{2\zeta}\right)\\
&= m\prod_{\zeta=t_i}^{t-1}\left(1+\frac{1}{2\zeta}\right)\\
&\approxeq m\exp\left(\sum_{\zeta=t_i}^{t-1}\frac{1}{2\zeta}\right)\\
&\approxeq m\exp\left(\int_{t_i}^{t}\frac{1}{2\zeta}d\zeta\right)\\
&= m\exp\left(\frac{1}{2}(\log t-\log t_i)\right)\\
&= m\sqrt{\frac{t}{t_i}}
\end{aligned}$$

s.t. error can be shown to be

$$\mathbb{E}[d_i(t)] = m\sqrt{\frac{t}{t_i}}\left(1+O\left(\frac{1}{t_i}\right)\right)$$

Heuristic largest degree

$$\begin{aligned}
d_{\max} &= \max_i d_i(t)\\
&\approxeq \max_i \mathbb{E}[d_i(t)] \quad t_i \text{ is minimal when } i=1\\
&= \mathbb{E}[d_1(t)]\\
&= m\sqrt{t}(1+O(1/t))
\end{aligned}$$

Heuristic degree distribution Let $N_k(t)$: number of vertices i with $d_i(t)=k$.

$$N_k(t) = \sum_{i=1}^{t+1}\mathbb{1}_{d_i=k}$$

We have:

$$\begin{aligned}
N_{\geq k}(t) &= \sum_{i=1}^{t+1}\mathbb{1}_{d_i\geq k} \quad \mathbb{E}[d_i] \approxeq m\sqrt{\frac{t}{t_i}} \approxeq m\sqrt{\frac{t}{i}}\\
&\approxeq \sum_{i=1}^{t+1}\mathbb{1}_{m\sqrt{\frac{t}{i}}\geq k}\\
&= \sum_{i=1}^{t+1}\mathbb{1}_{i\leq t\frac{m^2}{k^2}}\\
&= t\frac{m^2}{k^2}
\end{aligned}$$

Then, we have:

$$\frac{1}{t+1}N_{\geq k}(t) \to \frac{m^2}{k^2}$$

Thus, we have:

$$\begin{aligned}\frac{1}{t+1}N_k(t) &= \frac{1}{t+1}(N_{\geq k}(t) - N_{\geq k+1}(t+1)) \\ &\to \frac{m^2}{k^2} - \frac{m^2}{(k+1)^2} \\ &= \frac{m^2(2k+1)}{k^2(k+1)^2} \\ &\to \frac{2m^2}{k^3}\end{aligned}$$

i.e. power law degrees with exponent 3.

5.1 Properties

Theorem 5.1. $\exists$ r.v. $X_i,\ X \in \mathbb{R}_+$

1) $\frac{d_i(t)}{\sqrt{t}} \to X_i$ a.s. nodes

2) $\frac{N_k(t)}{t} \to \rho_k = \frac{2m(m+1)}{k(k+1)(k+2)}$ in probability Durrretl

3) $\frac{\max_i d_i(t)}{\sqrt{t}} \to X$ a.s. vH, Vol 1

4) $\text{diam}G(t) \sim \begin{cases} \log t & m = 1 \\ \frac{\log t}{\log\log t} & m \geq 2 \end{cases}$ vH, vol 2

5.1.1 $\frac{d_i(t)}{\sqrt{t}} \to X_i$ a.s.

Lemma 5.1. $M_t = \frac{d_i(t)}{\mathbb{E}[d_i(t)]}$ is a martingale.

Proof.

$$\begin{aligned}\mathbb{E}[d_i(t+1)] &= \mathbb{E}[\mathbb{E}[d_i(t+1)|d_i(t)]] \\ &= \mathbb{E}[d_i(t)\left(1+\frac{1}{2t}\right)] \\ &= \left(1+\frac{1}{2t}\right)\mathbb{E}[d_i(t)]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[M_{t+1}|G_t] &= \frac{1}{\left(1+\frac{1}{2t}\right)\mathbb{E}[d_i(t)]}\mathbb{E}[d_i(t+1)|d_i(t)] \\ &= \frac{1}{\left(1+\frac{1}{2t}\right)\mathbb{E}[d_i(t)]}d_i(t)\left(1+\frac{1}{2t}\right) \\ &= \frac{d_i(t)}{\mathbb{E}[d_i(t)]} = M_t\end{aligned}$$

□

Theorem 5.2. $\frac{d_i(t)}{\sqrt{t}} \to X_i$ a.s.

Proof 1. By the martingale convergence theorem, $M_t \to 1$ a.s. as $t \to \infty$. □

Proof 2. Calculate

1. $\mathbb{E}[N_k(t+1)] - \mathbb{E}[N_k(t)]$
2. $\rho_k = \lim_{t\to\infty} \frac{\mathbb{E}[N_k(t)]}{t}$
3. Prove $\frac{N_k(t)}{\mathbb{E}[N_k(t)]} \to 1$

□

5.1.2 $\frac{N_k(t)}{t} \to \rho_k = \frac{2m(m+1)}{k(k+1)(k+2)}$ in probability

Master equation $t \to t+1$ With probability $\frac{d_i(t)}{2m}$, $d_i(t) \to d_i(t)+1$ for each new edge

$$\mathbb{E}[N_k(t+1)|G_t] = N_k(t) + N_{k-1}(t)\frac{k-1}{2t} - N_k(t)\frac{k}{2t} + \delta_{k1} \quad \delta_{k1} : \text{new vertex}$$

General m

$$\mathbb{E}[N_k(t+1)\big|G_t] = N_k(t) + m\frac{k-1}{2mt}N_{k-1}(t) - m\frac{k}{2mt}N_k(t) + \delta_{km}$$

Let $Z_k(t) := \mathbb{E}[N_k(t)]$.

$$\boxed{Z_k(t+1) = Z_k(t) + \frac{1}{2t}\left[(k-1)Z_{k-1}(t) - kZ_k(t)\right] + \delta_{km}}$$

s.t. we fixed $m = 1$.

Theorem 5.3. $\frac{N_k(t)}{t} \to \rho_k = \frac{2m(m+1)}{k(k+1)(k+2)}$ in probability

Proof. Guess $Z_k(t) = tp_k$.

$$(t+1)p_k = tp_k + \frac{k-1}{2}p_{k-1} - \frac{k}{2}p_k + \delta_{km}$$

Then, we have:

$$\left(1+\frac{k}{2}\right)p_k = \frac{k-1}{2}p_{k-1} + \delta_{km}$$

i.e.

$$p_k = \frac{2}{k+2}\delta_{km} + \frac{k-1}{k+2}p_{k-1}$$

where $p_m = \frac{2}{m+2}$, $p_k = \frac{k-1}{k+2}p_{k-1}$ for $k > m$.

By induction, we have:

$$\boxed{p_k = \frac{2m(m+1)}{k(k+1)(k+2)}}$$

To prove convergence, we define $\Delta_k(t)$:

$$Z_k(t) = t(p_k + \Delta_k(t))$$

and show that $\Delta_k(t) \to 0$.

□

5.1.3 Largest degree

Theorem 5.4. $\frac{\max_i d_i}{\sqrt{t}} \to X$ a.s.

We will not prove this theorem. Instead, we prove bounds in probability.

Lemma 5.2. For $k = 1, 2, 3$, there exists $C_k < \infty$ s.t.

$$\mathbb{E}[d_i^k(t)] \leq C_k m^k \left(\frac{t}{t_i}\right)^{k/2}$$

Proof sketch. Note that we have:

$$\begin{aligned}
\mathbb{E}[d_i^2(t+1) \big| d_i(t)] &= d_i^2(t) + 2d_i(t)\,\mathbb{E}[X_i] + \mathbb{E}[X_i^2] \\
&\leq d_i^2(t) + \frac{d_i^2(t)}{t} + \frac{d_i(t)}{2t} + \frac{d_i^2(t)}{4t^2}
\end{aligned}$$

Then, we have:

$$\mathbb{E}[d_i^2(t+1)] \leq \left(1 + \frac{1}{t} + \frac{1}{4t^2}\right)\mathbb{E}[d_i^2(t)] + \cancel{\frac{1}{2t}\mathbb{E}[d_i(t)]}$$

i.e.

$$\begin{aligned}
\mathbb{E}[d_i^2(t)] &\leq m^2 \prod_{\zeta=t_i}^{t-1}\left(1 + \frac{1}{\zeta} + \frac{1}{4\zeta^2}\right) \\
&\leq m^2 \exp\left(\sum_{\zeta=t_i}^{t-1} \frac{1}{\zeta} + \frac{1}{4\zeta^2}\right) \\
&= m^2 \exp\left(\int_{t_i}^{t} \frac{1}{\zeta} d\zeta + O(1)\right) \\
&= C_2 m^2 \frac{t}{t_i}
\end{aligned}$$

$\square$

Corollary 5.1.

$$\Pr\left(\max_i d_i(t) \geq \lambda\sqrt{t}\right) = o\left(\frac{1}{\lambda^3}\right)$$

i.e. in particular, $\frac{d_{max}}{\sqrt{t}}$ is bounded in probability.

Proof.

$$\begin{aligned}
\Pr(\max_i d_i(t) \geq \lambda\sqrt{t}) &\leq \Pr\left(\sum_i d_i^3(t) \geq \lambda^3 t^{3/2}\right) \\
&\leq \frac{\sum_i \mathbb{E}[d_i^3(t)]}{\lambda^3 t^{3/2}} \\
&\leq \frac{m^3 C_3}{\lambda^3 t^{3/2}} \sum_i \frac{t^{3/2}}{t_i^{3/2}} \\
&= \frac{m^3 C_3}{\lambda^3} \left[1 + \sum_{i \geq 2} \frac{1}{(i-1)^{3/2}}\right] \\
&= o\left(\frac{1}{\lambda^3}\right)
\end{aligned}$$

□

5.1.4 Diameter of a graph

Theorem 5.5. $\mathrm{diam} G(t) \sim \begin{cases} \log t & m = 1 \\ \frac{\log t}{\log\log t} & m \geq 2 \end{cases}$

5.2 Global community structure

5.2.1 Stochastic Block Model (SBM)

$SBM(n, qB)$

- $B = (B\alpha\beta)$: symmetric, $k \times k$ matrix $B\alpha\beta \geq 0$
- q controls density

Model:

- assign color $\alpha(i) \in [k]$ i.i.d.
- connect i, j with probability $qB\alpha(i)\alpha(j)$

Properties

5.3 Local community structure

Main idea: Start with network with lots of local communities, then make it a small world.

Local community strategy Clustering coefficients

$$C_{\text{loc}} = \frac{1}{n}\sum_i \Pr(\text{two random friends of } i \text{ are friends}) = \frac{1}{n}\sum_i \frac{\text{number of } i \in \Delta}{\binom{d_i}{2}}$$
$$C_{\text{gl}} = \frac{\sum_i \text{number of } i \in \Delta}{\sum_i \binom{d_i}{2}}$$

Diameter abuse of notation

$$\text{diam} = \text{typical distance in } C_{\max}$$

For $G \sim G(n, p), CM(\underset{\sim}{d}), PA$

$$\text{diam}(G) \approxeq \log n$$
$$C_{\text{loc}}, C_{\text{gl}} \to 0 \quad \text{as } n \to \infty$$

Want: small diameter + large $C_{\text{loc}}, C_{\text{gl}}$

5.3.1 Household model

5.3.2 Watts, Strogatz '98 Small World

1. Start with cycles C_n degrees $= 2$, $C_{\text{loc}} = D$, diam $= n/2$
2. Connect all vertices with dist $\leq k$ degrees $= 2k$, $C_{\text{loc}} > D$, diam $= \frac{n}{2k}$
3. Rewire edges with probability p
    - randomly disattach one endpoint
    - connect that to a random $v \in [n]$

Heuristics if $1 < kp \leq \text{const}$; no rigorous proof

- long range edges behave like $G(n, kp/n)$

$$\text{diam} = \theta(\log n) = \theta\left(\frac{\log n}{kp}\right)$$

- short range edges shorten this to

$$\text{diam} = \theta\left(\frac{\log n}{k^2 p}\right)$$

5.3.3 Kleinberg '01-'02 Small World

Question: can a myopic agent find a target t in time $O(\log n)$?

Myopic agent

- can follow short and long edges
- sses one step ahead
- knows physical location of target and possible next moves

Question: how long does it take to reach target t?

Chapter 6

Directed models

Directed CM

Directed preferential attachment

Directed random graph $D(n, p)$

Part II

Processes and Algorithms

Chapter 7

Epidemics

Poisson Point Process (PPP) Define PPP on $\mathbb{R}_+$: random set of points $t_1 < t_2 < \cdots \in \mathbb{R}_+$

- t_k: kth click of a random clock
- $N(A) = \sum_i \mathbb{1}_{t_i \in A}$: number of clicks in A

We construct $PPP(\beta)$ on $\mathbb{R}_+$ as a limiting process. Define i.i.d. $\forall t \in \epsilon\mathbb{N}$, keep t with probability $\beta\epsilon$.
$N[a, b]$: number of clicks in $[a, b]$

$$\text{Bin}\left(\frac{b-a}{\epsilon}, \beta\epsilon\right) \to \text{Poi}(\beta(b-a)) \quad \text{as } \epsilon \to 0$$

Waiting time for first click is $T \sim \epsilon\text{Geom}(\beta\epsilon)$

$$\Pr(T \geq t) = (1-\beta\epsilon)^{t/\epsilon} \to e^{-\beta t}$$

7.1 Homogeneous mixing, differential equation

Homogeneous = complete graph. $S(t)$: fraction of susceptible individuals at time t, $I(t)$: infected, $R(t)$: recovered.

Note that we have $S(t) + I(t) + R(t) = 1$.

Define transmission rate β and recovery rate γ:

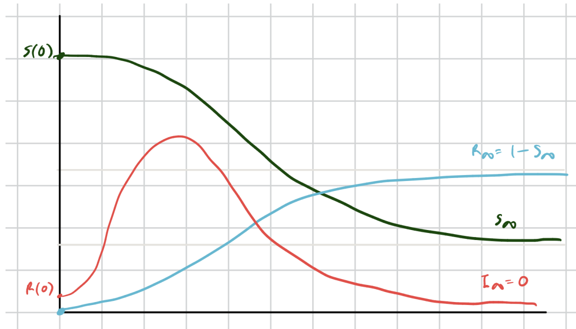
$$\frac{dS(t)}{dt} = -\beta S(t)I(t)$$
$$\frac{dI(t)}{dt} = \beta S(t)I(t) - \gamma I(t)$$
$$\frac{dR(t)}{dt} = \gamma I(t)$$

$R_0 = \frac{\beta}{\gamma}$: secondary infections by new infected individual before he recovers in $T \sim \text{Exp}(\gamma)$ time.

$$R_0 = \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|T]] = \mathbb{E}[\beta T] = \frac{\beta}{\gamma}$$

$R_0 \leq 1$: infection dies out exponentially fast.

$R_0 > 1$:



$$R_\infty + e^{-R_0 R_\infty} = 1$$

Differential equations

$$\frac{d \log S(t)}{dt} = \frac{1}{S(t)} \frac{dS(t)}{dt} = -\beta I(t)$$
$$\int_0^t \frac{d \log S(\tau)}{d\tau} d\tau = \log S(t) - \log S(0) := -\beta \int_0^t I(\tau) d\tau$$

i.e. $S(t) = S(0) e^{-\beta \int_0^t I(\tau) d\tau}$.

Verification:

$$\begin{aligned} \frac{dS(t)}{dt} &= S(0) e^{-\beta \int_0^t I(\tau) d\tau} \frac{d}{dt} - \beta \int_0^t I(\tau) d\tau \\ &= -\beta S(t) I(t) \end{aligned}$$

■

Assume the initial condition $R(0) = 0$. Note that $I(t) = \frac{1}{\gamma} \frac{dR}{dt}$

$$\int_0^t I(\tau) d\tau = \frac{1}{\gamma} \int_0^t \frac{dR}{d\tau} d\tau = \frac{1}{\gamma}(R(t) - R(0))$$

i.e. $S(t) = S(0) e^{-\frac{\beta}{\gamma} R(t)}$. ■

Denote $S(\infty) = \lim_{t \to \infty} S(t)$, $I(\infty) = \lim_{t \to \infty} I(t) = 0$, $R(\infty)$. Note that $S(\infty) + R(\infty) = 1$.

$$S(\infty) = S(0) e^{-\frac{\beta}{\gamma} R(\infty)} = 1 - R(\infty)$$

i.e. $R(\infty) = 1 - S(0) e^{-\frac{\beta}{\gamma} R(\infty)}$. ■

$$\begin{aligned} \frac{dS(t)}{dR(t)} &= -\frac{\beta S(t) I(t)}{\gamma I(t)} = -\frac{\beta}{\gamma} S(t) \quad R_0 = \frac{\beta}{\gamma} \\ &= -R_0 S(t) \end{aligned}$$

7.2 Homogeneous mixing, stochastic model

Consider $n+1$ people. At time $t=0$, 1 infected and n susceptible. If an individual i is infected:

- recovers at rate γ; i stays infected for $T_i \sim \text{Exp}(\gamma)$
- infects susceptibles at rate $\beta = c/n$

Defines random processes $S(t), I(t), R(t)$. Analysis of time dynamics is beyond scope. Focus on:

- probability of a large outbreak i.e. probability of $R(\infty) \geq \text{const } n$
- final size $R(\infty)$ conditioned on it being large
- conditions on β, γ for having a significant probability of a large outbreak

Theorem 7.1. If $R_0 = \frac{\beta}{\gamma} < 1$, infection dies out. If $R_0 > 1$

Infection digraph Predraw

$$T_i \sim \text{Exp}(\gamma) \quad \text{i.i.d. } \forall i$$
$$T_{ij} \sim \text{Exp}(\beta) \quad \text{i.i.d. } \forall i \neq j$$

Create a directed graph by putting an edge $i \to j$ if $T_{ij} < T_i$.

Note 1: If at time $t=0$, we start an epidemic at individual i, it reaches j at time

$$t_j = \min_{w: i \to j} \sum_{uv \in w} T_{uv}$$

Note 2: $R(\infty) = |C_t(1)|$

Forward BP What is the number X_1 of people infected by the first infected individual?

Conditioned on T_1, we have:

$$\begin{aligned}
\Pr(1 \text{ infects } i \big| T_1) &= \Pr(\beta\text{-clock clicks before } T_1) \\
&= \Pr_{T \sim \text{Exp}(\beta)}(T \leq T_1) \\
&= \int_0^{T_1} \beta e^{-\beta t} dt \\
&= -e^{-\beta t}\Big|_0^{T_1} = -e^{-\beta T_1} + 1 \\
&= 1 - e^{-\beta T_1} = p_{T_1}
\end{aligned}$$

Let X_1: number of people infected by 1

$$\begin{aligned}
\Pr(X_1 = k \big| T_1) &= \Pr(\text{Bin}(n, p_{T_1}) = k) \\
&= \binom{n}{k} p_{T_1}^k (1 - p_{T_1})^{n-k}
\end{aligned}$$

where the unconditional distribution is

$$\begin{aligned}\Pr(X_1 = k) &= \mathbb{E}_{T_1 \sim \text{Exp}(\gamma)}[p_{T_1}^k (1-p_{T_1})^{n-k}] \\ &= \binom{n}{k} \int_0^\infty \gamma e^{-\gamma t} p_t^k (1-p_t)^{n-k} dt\end{aligned}$$

Basic reproduction number Let $R_0 = \mathbb{E}[X_1]$: expected number of people infected by patient zero

$$\mathbb{E}[X_1 \big| T_1] = \mathbb{E}[\text{Bin}(n, p_{T_1})] = np_{T_1}$$

$$\begin{aligned}R_0 = \mathbb{E}[X_1] &= \mathbb{E}[\mathbb{E}[X_1 \big| T_1]] \\ &= n \int_0^\infty \gamma e^{-\gamma t}(1 - e^{-\beta t}) dt \\ &= n \left[\int_0^\infty \gamma e^{-\gamma t} dt - \int_0^\infty \gamma e^{-(\gamma+\beta)t} dt \right] \\ &= n \left(1 - \frac{\gamma}{\gamma + \beta} \right) \\ &= \frac{n\beta}{\gamma + \beta} \\ &= \frac{nc/n}{1 + c/n} \\ &= \frac{c}{1 + \frac{c}{n}} \to c\end{aligned}$$

More on forward BP Approximation: ignore that in later generations of the infection, we have less susceptible individuals. Thus, approximate $C_t(1)$ by T_{SIR}, where T_{SIR} is the BP with offspring distribution

$$p_k := \Pr(X = k) = \int_0^\infty \gamma e^{-\gamma t} \binom{n}{k} p_t^k (1-p_t)^{n-k}$$

Heuristic

$$\Pr(\text{epidemics survives}) = \Pr(|C_t(1)| \geq n^{1/3}) \approxeq \theta_{SIR}$$

We show that

$$\Pr(|T_{SIR}| = \infty) \to \theta_{SIR} = \begin{cases} 0 & c \leq 1 \\ 1 - \frac{1}{c} & c > 1 \end{cases}$$

Generating function and survival probability Generating function of a $\text{Bin}(n, p_T)$ is $(1 - p_T + p_T s)^n$.

For simplicity, assume $\gamma = 1$.

$$G_n(\lambda) = \int_0^\infty (1 - (1-\lambda)p_t)^n e^{-t} dt$$

Note that we have:

$$\begin{aligned}np_T &= n(1 - e^{-\beta T}) \\ &= n\left(1 - e^{-\frac{c}{n}T}\right) \\ &\approx n\frac{c}{n}T = cT\end{aligned}$$

and

$$\begin{aligned}(1-(1-\lambda)p_T)^n &= \exp\left(1-\frac{(1-\lambda)np_T}{n}\right)^n \\ &\approx \exp\left(1-\frac{(1-\lambda)cT}{n}\right)^n \\ &\to e^{-(1-\lambda)cT}\end{aligned}$$

i.e.

$$\begin{aligned}G_n(\lambda) &\to \int_0^\infty e^{-(1-\lambda)ct}e^{-t}dt \\ &= \int_0^\infty e^{-(1+c(1-\lambda))t}dt \\ &= \frac{1}{1+c(1-\lambda)} = G^c(\lambda)\end{aligned}$$

We solve $G^c(\eta) = \eta$ to get smaller one of $\eta = 1$ and $\eta = \frac{1}{c}$.

Concentration below criticality

$$\Pr\left(\sum_{i=1}^k X_i \geq kx\right) \leq \exp\left(-\frac{(x-c)^2}{2x(x+1)}k\right) \quad c \leq x \leq 1$$

7.3 Epidemic on configuration model graph

Idea: explore half-edges when β-clock clicks.

Given a $CM(\underset{\sim}{d})$. Define:

- $X(t)$: the number of active half-edges
- $S_d(t)$: the number of susceptible vertices of degree d
- $U(t)$: the number of unmatched half-edges

Rate of change:

$$\begin{aligned}\frac{dU}{dt} &= -2\beta X \\ \frac{dS_d}{dt} &= -\beta X\frac{dS_d}{U} \\ \frac{dX}{dt} &= -(\beta+\gamma)X - \beta X\frac{X}{U} + \beta X\sum_d \frac{dS_d}{U}(d-1)\end{aligned}$$

Chapter 8

Information cascades, influence maximization

Model: spread of information/innovation over social networks

8.1 Linear threshold model

A_{t+1}: inactive node v become active if $\sum_{u:\text{neighbor of } v} b_{uv} > \theta_v$.
θ_n: i.i.d. thresholds, uniform in [0, 1].

Influence of $S \subseteq V(G)$ $\quad I(S) = \mathbb{E}[|A_\infty \big| A_0 = S|]$

8.2 Independent cascade model (KKT '03)

Input: $D = (V, E)$, matrix $P = (p_{uv})_{uv \in E}$ of probabilities
Spreading process: every newly active node $v \in A_t \setminus A_{t-1}$ gets *one* chance p_{vu} to spread the information to $u \notin A_t$ such that $vu \in E$.

Theorem 8.1. Finding $\max_{|S|=k} I(S)$ is NP-complete in both models.

Definition 8.1. A monotonic increasing function $f : 2^V \to \mathbb{R}_+$, $f(S) \leq f(T)$ if $S \subseteq T$ is submodular iff

$$f(S \cup \{x\}) - f(S) \geq f(T \cup \{x\}) - f(T) \quad \forall S \subseteq T$$

i.e. diminishing returns.

Lemma 8.1. $I(S)$ is submodular for both models.

Chapter 9

Spectral methods

Motivation: define importance of webpage using only the WWW graph.

Algorithm 1: indegree. Importance of page i is equal to the number of pages linking to it, $\deg_i^{in}$.

Problem: one site, many votes d_i^{out}.

Algorithm 2: use scaled adjacency matrix $P_{ij} = \frac{1}{d_i^{out}} A_{ij} = D^{-1}A$. Importance of j is equal $\sum_i P_{ij}$. Or, iteratively, we have: $\pi(j) = \sum_i \pi(i) P_{ij}$. Scale w.r.t. fraction of votes each vertex is sending in general.

9.1 General Markov Chain

Fix state space $V = \{1, \dots, n\}$.
Define Δ: a set of probability distributions π on V, $M = (M_{uv})$: a stochastic matrix s.t. $M_{uv} \geq 0, \sum_v M_{uv} = 1 \quad \forall u \in V$.

Definition 9.1. A sequence of random variables $X_1, X_2, \dots \in V$ is called homogeneous (not t-dependent) Markov Chain with transition matrix M iff

$$\Pr(X_t = v_t \big| X_\zeta = v_\zeta \ \forall \zeta \leq t-1) = \Pr(X_t = v_t \big| X_{t-1} = v_{t-1}) = M_{v_{t-1}v_t}$$

Let $\mu_t(v) = \Pr(X_t = v)$. Given μ_0, we have:

$$\begin{aligned}\mu_t(v) &= \sum_u \Pr(X_{t-1} = u) \Pr(X_t = v \big| X_{t-1} = u) \\ &= \sum_u \mu_{t-1}(u) M_{uv} \\ &= (\mu_{t-1}M)_v = \dots = (\mu_0 M^t)_v\end{aligned}$$

i.e.

$$\boxed{\mu_t = \mu_0 M^t}$$

Definition 9.2. $\pi \in \Delta$ is called a stationary distribution of M iff $\pi = \pi M$.

Does π exist? Is it unique? Does $\mu_t \to \pi$?

Existance Let $v = (1, \dots, 1)^T$, $(Mv)_i = \sum_j M_{ij} v_j = 1 = v_i$. Thus, $\lambda = 1$ is always eigenvalue of the stochastic matrix M.

Since $\lambda = 1$ is eigenvalue of M, then $\exists$ left eigenvector $w \neq 0 : wM = \lambda w = w$.

Normalize w s.t. $\|w\|_1 = \sum_i |w_i| = 1$:

$$|w_j| = |(wM)_j| = \left| \sum_i w_i M_{ij} \right| \leq \sum_i |w_i| M_{ij}$$

Assume $\exists j : |w_j| < \sum_i |w_i| M_{ij}$. Then $\sum_j |w_j| < \sum_{i,j} |w_i| M_{ij} = \sum_i |w_i| = 1$. Contradiction.

Therefore, $|w_j| = (|w|M)_j$. Set $\pi_i = |w_i|$.

Theorem 9.1 (Perron Frobenius)**.**

- for all stochastic M, there exists $\pi : \pi M = \pi$
- for all stochastic M, $\psi M = \lambda \psi$ implies $|\lambda| \leq 1$
- if M is irreducible, there exists unique stationary distribution π and $\pi(v) > 0 \quad v \in V$
- if M is irreducible and aperiodic, $d(t) \to 0$ as $t \to \infty$ (convergence to stationarity)

Definition 9.3. D: finite digraph, C: SCC of D.

i) Period of C is equal GCD of all cycle lengths in C

ii) D is aperiodic iff for all its SCCs C, period of C is 1

Definition 9.4. A digraph is strongly connected and thus irreducible if you can get from any vertex to any other vertex via directed edges.

M is irreducible, aperiodic iff D_M is strongly connected, aperiodic.

Theorem 9.2.

(a) If M is irreducible (i.e. eigenvalue $\lambda = 1$ has multiplicity exactly 1) and aperiodic

9.2 Mixing time

Total Variation Distance

$$d_{TV}(\mu, \nu) = \max_{S \subseteq V} |\mu(S) - \nu(S)| = \frac{1}{2} \sum_{x \in V} |\mu(x) - \nu(x)|$$

$$\mu_{t,x}(v) = (M^t \mu_{0,x})(v) = M^t_{xv}$$

$$d(t) = \max_x d_{TV}(\mu_{t,x}, \pi)$$

$$\tilde{d}(t) = \max_{x,y} d_{TV}(\mu_{t,x}, \mu_{t,y})$$

Mixing time

$$\tau = \min\left\{t : \tilde{d} \le \frac{1}{e}\right\}$$

9.3 PageRank

Note that if $\mu_t \to \pi$, $\pi M = \pi$, $\pi(i)$ is still not a good measure of importance of i, since sinks will get high $\pi(i)$. Solution: restart with probability $\alpha \in [0,1]$:

$$\mu_t(v) = \alpha\frac{1}{n} + (1-\alpha)\sum_u \mu_{t-1}(u)P_{uv}$$

Thus, we have page rank of i equal $\pi(i)$, where π is a stationary distribution s.t. $\pi = \pi M$.

$$M_{uv} = \frac{\alpha}{n} + (1-\alpha)P_{uv} \tag{9.1}$$

i.e.

$$\boxed{\pi = \alpha\frac{1}{n}\mathbb{1}^{1\times n} + (1-\alpha)\pi P}$$

where P is a transition matrix of a digraph and $\pi \to \frac{1}{n}\mathbb{1}^{1\times n}$ as $\alpha \to 1$.

How fast does PageRank converge

Theorem 9.3.

$$\phi(t) \le (1-\alpha)^t \le e^{-\alpha t}$$

Proof. □

Random walks Given directed graph $D = (V, E)$ with adjacency matrix $A_{uv} = \mathbb{1}_{uv\in E}$, form random sequence $X_0, X_1, \dots \in V$ s.t. X_{t+1} conditioned on X_t is a uniformly random out-neighbor of X_t:s

$$\Pr(X_{t+1} = v \big| X_t = u) = P_{uv}$$

We study distribution of X_t as $t \to \infty$ and its convergence.

9.4 Spectrum, expansion, conductance

$G = (V, E)$: non-oriented, connected.

Transition matrix for random walk: $P = D^{-1}A$

$$P_{ij} = \begin{cases} \frac{1}{d_i}A_{ij} & \text{if } d_i > 0 \\ I_{ij} & \end{cases}$$

i.e. we add self-loops if there are no outgoing edges; where $D_{ij} = \delta_{ij}d_i$, $d_i = \sum_j A_{ij}$.

Transition matrix for lazy random walk that stays in the same place half of the time: $\tilde{P} = D^{-1/2}AD^{-1/2}$

$$\tilde{P}_{ij} = \begin{cases} \frac{1}{\sqrt{d_i}} A_{ij} \frac{1}{\sqrt{d_j}} & \text{if } d_i, d_j > 0 \\ I_{ij} & \end{cases}$$

where $\tilde{P}$ is symmetrized P.

P and $\tilde{P}$ are similar matrices since $\tilde{P} = D^{-1/2}AD^{-1/2} = D^{1/2}(D^{-1}A)D^{-1/2} = D^{1/2}PD^{-1/2}$.
Note that similar matrices have same eigenvalues.
Since $\tilde{P}$ is symmetric, all eigenvalues are real, all eigenvectors are orthogonal, $\lambda_2 = \max_{v \perp v_1} \frac{v^T \tilde{P} v}{v^T v}$. Given eigenbasis V, we have $\tilde{P}^t = V \Lambda^t V^T$.

Stationary distribution $\pi(i) = \frac{d_i}{2|E|}$.

Consider a transition matrix of a digraph $P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$.
Then, we have stationary distribution $\pi = \pi P$:

$$\pi(1) = \pi(1) + \frac{1}{2}\pi(2)$$
$$\dots$$
$$\pi(6) = 0$$

Applying PageRank $\pi = \alpha \frac{1}{n} \mathbb{1}^{n \times 1} + (1-\alpha)\pi P$, we have:

$$\pi(1) = \frac{\alpha}{n} + (1-\alpha)\left[\pi(1) + \frac{1}{2}\pi(2)\right]$$
$$\dots$$
$$\pi(6) = \frac{\alpha}{n}$$

where $n = 6$.

Spectrum of P $\quad P\psi_i = \lambda\psi_i \quad \lambda_i = 1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq -1$.

Key: $v = (1, \dots, 1)^T$ is right eigenvector of P for eigenvalue $\lambda 1$: $v = Pv$.

Spectrum of $\tilde{P}$ $\quad \psi_1 : \psi_1(i) = \sqrt{\pi(i)}$ is eigenvector for eigenvalue λ_1.

Orthogonal eigenvectors

$$(\psi_i, \psi_j)_D = \psi_i^T D \psi_j = 0 \quad \forall i \neq j$$

Theorem 9.4 (Spectrum relation to mixing)**.** For the lazy chain with transition matrix $M = \frac{P+I}{2}$ i.e. don't move with probability 1/2, we have:

$$\lambda_2^t \leq d(t) \leq \frac{\lambda_2^t}{\sqrt{\min_i \pi(i)}}$$

Proof. Note that if G has $k \geq 2$ connected components, then eigenvalue 1 of P has degeneracy at least k. Let $C_1 = (V_1, E_1), \ldots, C_k = (V_k, E_k)$ be components of G. We know that P makes no transitions between different components, so $1 = \sum_v P_{uv} = \sum_{v \in V_1} P_{uv}$ yielding that the vector $\psi^{(i)} = \mathbb{1}_{v_i}$ is eigenvector with eigenvalue 1, giving k lienarly independent eigenvectors. ■

By definition of $\mu_{t,x}(v) = P_{xv}^t$.

Choosing any $x, y \in V$ such that they belong to difference clusters (connected compoents), we have:

$$\begin{aligned}
d_{TV}(\mu_{t,x}, \mu_{t,y}) &= \frac{1}{2} \sum_v |\mu_{t,x}(v) - \mu_{t,y}(v)| \\
&= \frac{1}{2} \sum_{v \in V_x} |\mu_{t,x}(v)| + \frac{1}{2} \sum_{v \in V_y} |\mu_{t,y}(v)| \\
&= \frac{1}{2} \sum_{v \in V_x} \mu_{t,x}(v) + \frac{1}{2} \sum_{v \in V_y} \mu_{t,y}(v) \\
&= 1
\end{aligned}$$

Thus, we have $\tilde{d}(t) = \max_{x,y \in V} d_{TV}(\mu_{t,x} - \mu_{t,y}) = 1$. ■

Thus, we have $\tilde{\lambda}_2 \leq \tilde{d}(t) = 1$. □

Note that $\lambda_2 \tilde{P} = \lambda_2 D^{1/2} \psi_2$ since

□

Theorem 9.5.

$$\frac{\phi^2}{8} \leq 1 - \lambda_2 \leq \phi$$

where $1 - \lambda_2$ is a spectral gap.

Theorem 9.6 (Expanders have small diameter)**.** Given $G = (V, E)$.

Define $\alpha = \min_S \frac{|dS|}{|S|} = \min_{0 < |S| \leq \frac{n}{2}} \frac{\text{cut}(S, S^C)}{|S|}$.

Then, $\exists C_{d_{\max},\alpha} : \text{diam}(G) \leq 2 + 2 \log_{C_{d_{\max},\alpha}} \frac{n}{2}$.

Proof.

$$d_{\max} |S_{i+1} \setminus S_i| \geq |dS_i| \geq \alpha |S_i|$$

i.e. $|S_{i+1} \setminus S_i| \geq \frac{\alpha}{d_{\max}} |S_i|$.

Consider $t_x = \min \left\{ t : |S_t| < \frac{n}{2} \right\}$. Then, we have $|S_t| \leq \frac{n}{2}$ $t = 1, \ldots, t_x - 1$ and $|S_t| = |S_{t-1}| + |S_t \setminus S_{t-1}| \geq |S_{t-1}| \left(1 + \frac{\alpha}{d_{\max}}\right) = |S_0| \left(1 + \frac{\alpha}{d_{\max}}\right)^t = C^t$ where $|S_0| = |\{x\}| = 1$.

Thus, we have $\frac{n}{2} \geq |S_{t_x - 1}| \geq C^{t_x - 1}$, and $1 + \log_C \frac{n}{2} \geq t_x$.

Similarly, same holds for any other vertex $y \in V$ and t_y.

Note that we have $\text{diam}G = max_{x,y \in V} \text{dist}(x, y) \leq t_x + t_y \leq 2 + 2 \log_C \frac{n}{2}$. □

Theorem 9.7 (Your friends have more friends than you)**.** For any graph G, we have:

$$d^*(G) \geq \bar{d}(G)$$

where $d^*(G)$: average degree of neighbors.

Proof. Let A be adjacency matrix of G. Then, we have $d_i = \sum_j A_{ij}$:

$$d_i^* = \frac{1}{d_i} \sum_j A_{ij} d_j$$

Thus, we have:

$$\begin{aligned}
d^*(G) &= \frac{1}{n} \sum_i d_i^* \\
&= \frac{1}{n} \sum_i \frac{1}{d_i} \sum_j A_{ij} d_j \\
&= \frac{1}{n} \sum_{ij} A_{ij} \frac{d_j}{d_i} \quad (d_i - d_j)^2 \geq 0;\ \frac{d_i}{d_j} + \frac{d_j}{d_i} \geq 2\ \forall i, j \\
&= \frac{1}{n} \sum_{i \leq j} A_{ij} \left(\frac{d_i}{d_j} + \frac{d_j}{d_i} \right) \\
&\geq \frac{1}{n} 2 \sum_{i \leq j} A_{ij} \quad A_{ij} = A_{ji} \\
&= \bar{d}(G)
\end{aligned}$$

□

9.5 Spectral clustering (K-Means)

- Solve $P\psi_2 = \lambda_2 \psi_2$ where λ_2: second largest eigenvalue.
- Output: $S = \{i \in V : \psi_2(i) \leq \alpha\}$.
- Either (a) optimize α or (b) sign cut $\alpha = 0$

Theorem 9.8.

$$1 - \lambda_2 \leq \phi \leq \phi(S)$$

Problem 9.1. Two planted communities

$$1 - \lambda = 1 - \frac{n_1 n_2 (ab - \epsilon^2)}{D_1 D_2}$$

Problem 9.2. Laplacian on graphs with several connected components

$$(D - A)\psi = L\psi = \Lambda D\psi$$

where $\Lambda = 1 - \lambda$.

Note that the Laplacian is PSD:

$$\psi^T(D - A)\psi = \frac{1}{2}\sum_{x,y\in V}(\psi_x - \psi_y)^2 A_{xy} \geq 0$$

Chapter 10

Reference

Binomial theorem

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

Partial sums

$$S_n = \frac{a(1-r)^n}{1-r}$$

$$\sum_{k\geq 0} kr^k = \frac{r}{(1-r)^2} \quad r \in (0,1)$$

PGFs, MGFs

Taylor expansion Taylor expansion of a function f around a point a up to order n for some ξ between a and x:

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + R_n(x)$$

Lagrange remainder

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1}$$

Integral form

$$R_n(x) = \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt$$

Asymptotic form

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + o((x-a)^n) \quad \text{given } x \to a$$

Fundamental Theorem of Calculus

$$\frac{d}{dx}\int_a^x f(t)dt = f(x)$$

$$\int_a^b f'(x)dx = f(b) - f(a)$$