Probabilistic Aspects of Computer Science
Random graphs

Anne Bouillard

December 7, 2020
Galton-Watson branching processes

The Galton-Watson branching process was initially introduced to study the extinction of family names in the Victorian England.

- $X_0 = 1$ (root, depth 0)
- $X_n$ number of nodes at depth $n$ (population of the $n$-th generation)
- $Z_i^{(n)}$ number of children of node $i$ of the $n$-th generation. The $(Z_i^{(n)})_{i,n}$ are i.i.d.

What is the probability that the tree is finite?
$Z$ is a r.v. with the same distribution as $Z_i^{(n)}$.

**Theorem**

Let $p_e$ be the extinction probability of the Galton-Watson process.

1. *If* $P(Z > 1) > 0$ *and* $E[Z] \leq 1$ *then* $p_e = 1$;
2. *If* $P(Z > 1) = 0$ *and* $E[Z] = 1$, *then* $p_e = 0$;
3. *If* $E[Z] > 1$, *then* $p_e = \beta < 1$. 
Distribution of the population of the $n$-the generation

$$X_{n+1} = \sum_{i=1}^{X_n} Z_i^{(n)}.$$

**Generating functions:**
- $g(s) = \mathbb{E}[s^Z]$ the generating function of $Z$
- $\phi_n = \mathbb{E}[s^{X_n}]$ that of $X_n$.

**Lemma**

$$\phi_{n+1} = g_Z(\phi_n).$$

- Wald equality: $\phi_{n+1} = \phi_n \circ g_Z$.
- $\phi_{n+1} = \phi_0 \circ g_Z \circ \cdots \circ g_Z = \phi_0 \circ g_Z^{n+1}$.
- $\mathbb{P}(X_0 = 1) = 1$, so $\phi_0(s) = s$ and $\phi_{n+1} = g_Z^{n+1}$. 

A. Bouillard PACS December 7, 2020 4 / 27
Extinction probability as a fix-point equation

- \( p_e = \mathbb{P}(\exists n \in \mathbb{N}, \ X_n = 0) = \mathbb{P}(\bigcup_{n \in \mathbb{N}} \{X_n = 0\}) \) the extinction probability of the process
- \( \{X_n = 0\} \subseteq \{X_{n+1} = 0\}, \text{ so } p_e = \lim_{n \to \infty} \mathbb{P}(X_n = 0). \)

Lemma

\[ p_e = g_Z(p_e). \]

- We know that \( \phi_{n+1}(0) = g_Z(\phi_n(0)) \)
- we also have \( \phi_{n+1}(0) = \mathbb{P}(X_{n+1} = 0) \) and \( \phi_n(0) = \mathbb{P}(X_n = 0) \)
- Then, by continuity (\( g_Z \) is continuous on \( [0, 1] \)), \( p_e = g_Z(p_e). \)
Extinction probability as a fix-point equation (2)

**Theorem (fixed point)**

Consider the equation \( p = g(p) \) where \( g \) is the generating function of a random variable \( X \).

1. \( g \) is non-decreasing and convex on \([0, 1]\). Moreover, if \( \Pr(X = 0) < 1 \), then \( g \) is strictly increasing, and if \( \Pr(X \leq 1) < 1 \), then \( g \) is strictly convex.

2. If \( \Pr(X < 1) < 1 \), and if \( \mathbb{E}[X] \leq 1 \), then the equation \( x = g(x) \) has a unique solution in \([0, 1]\), \( x = 1 \). If \( \mathbb{E}[X] > 1 \), then the equation \( x = g(x) \) has two solutions, in \([0, 1]\), \( x = 1 \) and \( \beta \in [0, 1] \).

- \( g_Z(s) = \sum_{n \in \mathbb{N}} \Pr(Z = n)s^n \) is non-decreasing and strictly increasing if \( \Pr(Z = 0) < 1 \).
- \( g'_Z(s) = \sum_{n \in \mathbb{N}} \Pr(Z = n + 1)s^n \) is non-decreasing and strictly increasing if \( \Pr(Z \leq 1) < 1 \), so \( g_Z \) is convex and strictly convex if \( \Pr(Z \leq 1) < 1 \).
Extinction probability as a fix-point equation (3)

- $x = 1$ is trivially a solution.
- Now, we use the convexity of $g_Z$.
- If $\mathbb{E}[X] \leq 1$, then $g'_Z(1) \leq 1$ and, as the function is convex, $\forall x < 1$, $g'_Z(x) \leq 1$ and $g_Z(x) > x$.
- If $\mathbb{E}[X] > 1$, on an interval $[1 - \epsilon, 1[, g_Z(x) < x$. But $g_Z(0) \geq 0$, so there exists $\beta$ such that $\beta = g_Z(\beta)$.
Extinction probability

**Theorem**

Let $p_e$ be the extinction probability of the Galton-Watson process.

1. If $P(Z > 1) > 0$ and $E[Z] \leq 1$ then $p_e = 1$;
2. If $P(Z > 1) = 0$ and $E[Z] = 1$, then $p_e = 0$;
3. If $E[Z] > 1$, then $p_e = \beta < 1$.

- $x_n = P(X_n = 0)$
- $x_0 = 0$, so $\beta - x_0 \geq 0$
- if $x_n \leq \beta$, then as $g_Z$ is non-decreasing, $x_{n+1} = g_Z(x_n) \leq g_Z(\beta) = \beta$
- So $p_e \leq \beta$ and finally $p_e = \beta$
Emergence of the giant component

Size of the connected components.

- $C_1$ size of the largest connected component
- $C_2$ size of the second largest connected component

With $p = c/n$,

**Theorem**

(i) *(sub-critical regime)* If $c < 1$, then there exists $a > 0$ such that

$$\lim_{n \to \infty} P(|C_1| \leq a \ln n) = 1.$$  

(ii) *(critical regime)* If $c = 1$, the there exists $\kappa > 0$ such that for all $a > 0$,

$$\lim_{n \to \infty} P(|C_1| \geq an^{2/3}) \leq \frac{\kappa}{a^2}.$$  

(iii) *(super-critical regime)* If $c > 1$, there exists a unique $p_e \in ]0, 1[$ and there exists $a > 0$ such that for all $\delta > 0$,

$$\lim_{n \to \infty} P\left(|\frac{|C_1|}{n} - (1 - p_e)| \leq \delta \text{ and } |C_2| \leq a' \ln n\right) = 1.$$
Emergence of the giant component

$p = 0.4000$

$p = 0.8000$

$p = 1.0000$

$p = 1.5000$
Link with Galton-Watson processes

- $u$ a vertex, $C(u)$ the connected component it belongs to.
- Breadth-first search from $u$ of the connected component of $u$

\[ Z_0 \sim \text{Bin}(n - 1, p) \]
\[ Z_1 \sim \text{Bin}(n - Z_0, p) \]
\[ Z_2 \sim \text{Bin}(n - Z_0 - Z_1, p) \]
\[ Z_3 \sim \text{Bin}(n - Z_0 - Z_1 - Z_2, p) \]
\[ \ldots \]

Almost like $Z \sim \text{Bin}(n, p)$
Analysis of $C(u)$

Vertices at time $t$ can be
- live (queued vertices),
- neutral (vertices not discovered yet)
- or dead (popped vertices)

The BFS from $u$ is done the following way:
- Initially (at time $t = 0$), every vertex is neutral except $u$, which is live: $L(0) = 1$, $D(0) = 0$ and $N(0) = n - 1$
- At each time $t$, we take one live vertex $w$ in the queue, pop it and queue all its neighbors that are still neutral. Then those vertices become live and $w$ becomes dead.
- The procedure ends when the queue is empty, and the dead vertices correspond to $C(u)$. Let $Z(t)$ be the number of vertices added in the queue at time $t$.

$$L(t) = L(t - 1) - 1 + Z(t), \quad N(t) = N(t - 1) - Z(t) \quad \text{and} \quad D(t) = t.$$ 

In other words, $N(t) = n - t - L(t)$ and $Z(t)$ is found by checking the adjacency between one vertex and $N(t)$ vertices, that is $Z(t) \sim \text{Bin}(N(t - 1), p) = \text{Bin}(n - t + 1 - L(t - 1), p)$. 

Extinction of a branching process

- Let $T$ be the first time when there is no live vertex.
- $T$ is the smallest integer $t$ such that $L(t) = 0$. The process stops at time $T$ and $T$ is the size of the component.
- $L(0) = 1$ and for all $t > 0$, $L(t) = L(t - 1) + Z(t) - 1$.
- The process stops when $L(t) = 0$, but the variables $L(t)$ can still be defined after the process stops.
- In that case, for all $t$,

$$L(t) = L(0) + \sum_{s=1}^{t} Z(s) - t = \sum_{s=1}^{t} Z(s) - t + 1$$

- Similarly for branching processes. If the branching process does not stop, then $T = \infty$. 
Emergence of the giant component

Comparison of the graph process with the binomial process

A binomial process is when $Z \sim \text{Bin}(n, p)$. Here, contrary to the graph process, $n$ does not change with the size of the branching.

- $T_{n,p}^{\text{bin}}$: size of the binomial branching process with parameters $n$ and $p$
- $T_{n,p}^{\text{gr}}$: size of the graph process from a vertex in $\mathcal{G}(n, p)$.

Lemma

For any $k \in \mathbb{N}$,

$$\mathbb{P}(T_{n-k,p}^{\text{bin}} \geq k) \leq \mathbb{P}(T_{n,p}^{\text{gr}} \geq k) \leq \mathbb{P}(T_{n-1,p}^{\text{bin}} \geq k) \leq \mathbb{P}(T_{n,p}^{\text{bin}} \geq k).$$
Comparison of the graph process with the binomial process: proof
Sub-critical regime: \( c < 1 \)

\[
\Pr(T_{n,p}^{gr} > k) \leq \Pr(T_{n,p}^{bin} > k) \\
\leq \Pr(\text{Bin}(nk, p) \geq k) \\
\leq \Pr\left(\text{Bin}(nk, p) \geq kc(1 + \left(1 - \frac{1}{c}\right))\right) \\
\leq e^{-\frac{kc}{3} \left(\frac{c-1}{c}\right)^2} \text{ Chernoff bound}
\]
The sub-critical regime: $c < 1$

\[
\mathbb{P}(T_{n,p}^{gr} \geq k) \leq e^{-\frac{k c}{3} \left(\frac{c-1}{c}\right)^2}
\]

Set $k = a \ln n$, then

\[
\mathbb{P}(T_{n,p}^{gr} \geq a \ln n) \leq n^{-a \frac{(c-1)^2}{3c}}.
\]

If we choose $a = \frac{4c}{(c-1)^2}$, then we obtain $\mathbb{P}(|C(u)| \geq a \ln n) \leq n^{-4/3}$ and

\[
\mathbb{P}(C_1 \geq a \ln n) \leq \sum_u \mathbb{P}(|C(u)| \geq a \ln n) \leq n^{-1/3} \xrightarrow{n \to \infty} 0.
\]
There are small and giant components only

Let $k^- = a' \ln n$ and $k^+ = n^{2/3}$.

**Lemma**

*For all vertex $v$, with high probability,*

- either the branching process from $v$ stops before $k^-$ steps (i.e. $|C(v)| \leq k^-$);
- or $\forall k, k^- \leq k \leq k^+$, there are at least $\frac{(c-1)k}{2}$ live vertices ($L_v(k) \geq \frac{(c-1)k}{2}$).

We call *bad vertex* a vertex satisfying none of these properties.

Let $v$ be a vertex. Either the branching process stops in less than $k^-$ steps, or in more than $k^-$ steps.

$v$ is a bad vertex in the second case and if there exists $k \in ]k^-, k^+]$ such that $L_v(k) < \frac{(c-1)k}{2}$,
that is the number of visited vertices at step $k$ is strictly less than $k + \frac{(c-1)k}{2} = \frac{(c+1)k}{2}$. 
Proof of the lemma

Let $B(v, k)$ the event "$v$ is a bad vertex at step $k$" (there are less than $\frac{(c+1)k}{2}$ visited vertices), for $k \in [k^-, k^+]$.

\[
P(B(v, k)) \leq P\left(\sum_{i=1}^{k} B_{\text{in}}(n - \frac{(c+1)k}{2}, \frac{c}{n}) \leq \frac{(c+1)k}{2} - 1\right)
\]
\[
\leq P(B_{\text{in}}(k(n - \frac{(c+1)k}{2}), \frac{c}{n}) \leq \frac{(c+1)k}{2} - 1)
\]
\[
\leq P(B_{\text{in}}(k(n - \frac{(c+1)k^+}{2}), \frac{c}{n}) \leq \frac{(c+1)k}{2})
\]
We use Chernoff bounds: \( \mathbb{E}[\text{Bin}(k(n -(c+1)k^+/2), \frac{c}{n})] = ck(1 - \frac{(c+1)k^+}{2n}) \) and choose \( \delta \) such that \((1 - \delta)ck(1 - \frac{(c+1)k^+}{2n}) = \frac{(c+1)k}{2} \):

\[
\delta = 1 - \frac{(c+1)}{c(2 - (c+1)k^+/n)} \xrightarrow{n \to \infty} 1 - \frac{(c+1)}{2c} = \frac{c-1}{2c}.
\]

So,

\[
P(B(v, k)) \xrightarrow{n \to \infty} p \leq \exp\left(-\frac{(c-1)^2}{8c}k\right)
\]

and more precisely after computations,

\[
P(B(v, k)) \leq \exp\left(-\frac{(c-1)^2}{8c} + O(n^{-1/3})k\right)
\]
The probability that $v$ is a bad vertex is bounded by

$$P\left(\bigcup_{k=k^-}^{k^+} B(v, k)\right) \leq \sum_{k=k^-}^{k^+} e^{-\left(\frac{(c-1)^2}{8c} + O(n^{-1/3})\right)k}$$

$$\leq n^{2/3} e^{-\left(\frac{(c-1)^2}{8c} + O(n^{-1/3})\right)k^-}$$

$$\leq n^{2/3} e^{-\left(\frac{(c-1)^2}{8c} + O(n^{-1/3})\right)a' \ln n} = n^{2/3} n^{-\left(\frac{(c-1)^2}{8c} a' + O(n^{-1/3})\right)}.$$

With $a' = \frac{16c}{(c-1)^2}$, this probability is less than $n^{-4/3}$ and the probability that there exists a bad vertex is less than $n^{-1/3}$.

We call a *small vertex* a vertex satisfying the first property and a *large vertex* a vertex satisfying the second.
There is at most one giant component

- $u$ and $v$ are two large vertices.
- $U(u)$ and $U(v)$ are the sets of live vertices after $k^+$ steps of the branching processes from $u$ and from $v$
- $|U(u)| \geq \frac{(c-1)k^+}{2}, |U(v)| \geq \frac{(c-1)k^+}{2}$ and assume that $U(u) \cap U(v) = \emptyset$

\[
P(C(u) \neq C(v)) \leq P(\text{there is no arc between } U(u) \text{ and } U(v)) \leq (1 - p)|U(u)| \cdot |U(v)| \leq (1 - p)\left(\frac{(c-1)k^+}{2}\right)^2 \leq e^{-p\left(\frac{(c-1)k^+}{2}\right)^2} \leq e^{-\frac{(c-1)^2c}{4}n^{1/3}} = o(n^{-2})
\]

Consequently, $P(\text{there are several large components}) = o(1)$. 
Number of small vertices

- Let $N_s$ the number of small vertices.
- Let $T^-$ be the size of a Galton-Watson branching process with offspring law $\text{Bin}(n - k^-, c/n)$.
- $T^+$ be the size of a Galton-Watson branching process with offspring law $\text{Bin}(n, c/n)$.
- $T = |C(u)|$

\[
P(T^+ \leq k^-) \leq P(T \leq k^-) \leq P(T^- \leq k^-).
\]

But, when $k$ is fixed, $P(T^- \leq k^-) \xrightarrow{n \to \infty} P(T_c^{\text{poi}} \leq k)$ and $P(T^+ \leq k^-) \xrightarrow{n \to \infty} P(T_c^{\text{poi}} \leq k)$.

Now, when $k$ grows to $\infty$, $P(T_c^{\text{poi}} \leq k) \xrightarrow{k \to \infty} p_e$. Then

\[
E[N_s] = (p_e + o(1))n.
\]
The large component is a giant component

Let $S_u$ the r.v. equal to 1 if $u$ is a small vertex and 0 otherwise. We have $N_s = \sum_u S_u$

\[
E[N_s^2] = \sum_u E[S_u] + \sum_{u\neq v} E[S_uS_v] = E[N_s] + \sum_v P(S_v = 1) \sum_{u\neq v} P(S_u = 1 | S_v = 1).
\]

But for all $v$,

\[
\sum_{u\neq v} P(S_u = 1 | S_v = 1) = \sum_{u\neq v, u \in C(v)} P(S_u = 1 | S_v = 1) + \sum_{u \notin C(v)} P(S_u = 1 | S_v = 1) \leq k^- + (p_e + o(1))n = (p_e + o(1))n.
\]
So
\[
\text{Var}(N_s) \leq E[N_s] + n^2(p_e + o(1))^2 - E[N_s]^2 \\
\leq E[N_s] + o(E[N_s]^2)
\]
Then
\[
P(|N_s - E[N_s]| \geq \delta E[N_s]) \leq \frac{\text{Var}(N_s)}{\delta^2 E[N_s]^2} = \frac{1}{\delta^2} \left( \frac{1}{E[N_s]} + o(1) \right) = o(1).
\]
Application to the epidemic model SIR

Random graphs can be seen as a model for epidemic processes. Consider the Reed-Frost model: consider a population of $n$ individuals.

- At time 0, a single individual is infected.
- When an individual is infected, it is infectious during one time step, and after, it is removed (dead or immunized...).
- While infectious, it can infect every other healthy individual with probability $p$, and independently of the other infections.

More formally,

- let $Z_u(t) \in \{S, I, R\}$ be the state (susceptible, infected, removed) of vertex $u$ at time $t$,
- $Z(t) = (Z_u(t))_u$ the global state at time $t$.
- We denote by $S(z)$, $I(z)$ and $R(z)$ the number of susceptible, infected, removed vertices in state $z$. 
Application to the epidemic model SIR

The process can be modeled by a Markov chain:

\[
P(Z(t + 1) = z' \mid Z(t) = z) = \begin{pmatrix} S(z) \\ I(z') \end{pmatrix} (1 - p)^{I(z)S(z')} [1 - (1 - p)^{I(z)}]^{I(z')}
\]

if \( z_i \in \{I, R\} \Rightarrow z_i' = R \) and \( z_i = S \Rightarrow z_i' \in \{S, I\} \).

This model can also be studied using Erdős-Rényi graphs:

- if \( u \) is originally infected, then the size of the epidemic is the size of the connected component \( C(u) \) in \( G(n, p) \).
- if \( p = c/n \) with \( c < 1 \), then with high probability, the number of infected individuals is then \( O(\ln n) \),
- if \( c > 1 \), then with probability \((1 - p_e) + o(1)\), a proportion \((1 - p_e) + o(1)\) is infected. The average size of the epidemic is \((1 - p_e)^2 n + a' p_e \ln n \sim (1 - p_e)^2 n\).