Lecture 13: Random Graph (Chapter 8 of Textbook B)

Jinwoo Shin

AI503: Mathematics for AI

Roadmap

KAIST AI
Graduate School of AI

(1) The $G(n, p)$ Model

(2) Phase Transition

(3) Giant Component

(4) Cycles and Full Connectivity

(5) Phase Transition for Increasing Properties

(6) Branching Processes

Roadmap

KAIST AI
Graduate School of AI

(1) The $G(n, p)$ Model

(2) Phase Transition

(3) Giant Component

(4) Cycles and Full Connectivity

(5) Phase Transition for Increasing Properties

(6) Branching Processes

The $G(n, p)$ Model

KAIST AI
Graduate School of AI

- One of the most basic model of a random graph is the $G(n, p)$ model.
- The $G(n, p)$ model has two parameters:
  - n is the number of vertices of the graph
  - p is the edge probability
- For each pair of distinct vertices v and w, p is the probability that the edge (v, w) is present.
- The graph-valued random variable with these parameters is denoted by $G(n, p)$.
- When we refer to "the graph $G(n, p)$", we mean one realization of the random variable.

Degree Distribution

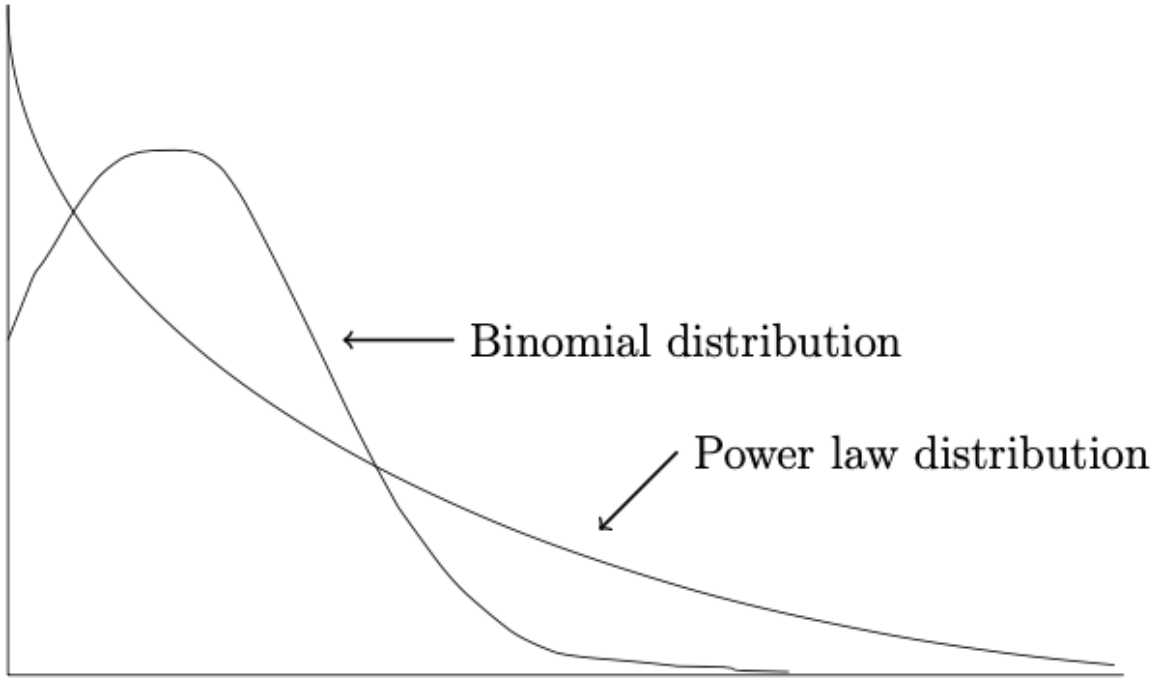
- One of the simplest quantities to observe in a real graph is the number of vertices of given degree, called the vertex degree distribution.
- Since the degree of each vertex is the sum of n independent random variables, which results in a binomial distribution.

$$Prob(\text{degree of vertex} = k\ ) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \approx \binom{n}{k} p^k (1-p)^{n-k}$$

- The binomial distribution falls off exponentially fast as one moves away from the mean.

Degree Distribution

- However, the degree distribution of the graphs that appear in many applications do not exhibit sharp drops.
- Rather, many large graphs that arise in various applications appear to have power law degree distributions.



Degree Distribution

- The following theorem states that the degree distribution of the random graph $G(n, p)$ is tightly concentrated about its expected value, i.e. the probability that the degree of a vertex differs from its expected degree by more than $\lambda\sqrt{np}$, drops off exponentially fast with λ.

Theorem.

Let v be a vertex of the random graph $G(n, p)$. Let $\alpha \in (0, \sqrt{np})$. Then,

$$Pr(|np - deg(v)| \geq \alpha\sqrt{np}) \leq 3e^{-\alpha^2/8}.$$

Proof. See page 249 of Textbook B.

Degree Distribution

KAIST AI
Graduate School of AI

Corollary

Suppose ε is a positive constant. If $p \geq \frac{9 \ln n}{n\varepsilon^2}$, then almost surely every vertex has degree in the range $(1-\varepsilon)np$ to $(1+\varepsilon)np$.

Proof. Apply the previous theorem with $\alpha = \varepsilon\sqrt{np}$ to get that the probability that an individual vertex has degree outside the range $[(1-\varepsilon)np, (1+\varepsilon)np]$ is at most $3e^{-\varepsilon^2 np/8}$. By the union bound, the probability that some bertex has degree ouside this range is at most $3ne^{-\varepsilon^2 np/8}$. For this to be $o(1)$, it suffices for $p \geq \frac{9 \ln n}{n\varepsilon^2}$.

Existence of Triangles in $G(n, d/n)$

Question: What is the expected number of triangles in $G(n, \frac{d}{n})$, when d is a constant?

- Let Δ_{ijk} be the indicator variable for the triangle with vertices i, j and k being present. (i.e., all three edges $(i, j), (j, k)$ and (k, i) being present.
- Then the number of triangle is $x = \sum_{ijk} \Delta_{ijk}$.
- Thus, the expected number of triangles is

$$E(x) = E\left(\sum_{ijk} \Delta_{ijk}\right) = \sum_{ijk} E\left(\Delta_{ijk}\right) = \binom{n}{3}\left(\frac{d}{n}\right)^3 \approx \frac{d^3}{6}.$$

- However, this does not mean that a graph has a triangle with high probability.

Roadmap

KAIST AI
Graduate School of AI

(1) The $G(n, p)$ Model

(2) Phase Transition

(3) Giant Component

(4) Cycles and Full Connectivity

(5) Phase Transition for Increasing Properties

(6) Branching Processes

Phase Transition

- Many properties of random graphs undergo structural changes as the edge probability passes some threshold value.
- This phenomenon is similar to the abrupt phase transitions in physics, as the temperature or pressure increases.

Probability	Transition
$p = o(\frac{1}{n})$	Forest of trees, no component of size greater than $O(\log n)$
$p = \frac{d}{n}, d < 1$	Cycles appear, no component of size greater than $O(\log n)$
$p = \frac{d}{n}, d = 1$	Components of size $O(n^{\frac{2}{3}})$
$p = \frac{d}{n}, d > 1$	Giant component plus $O(\log n)$ components
$p = \frac{1}{2}\frac{\ln n}{n}$	Giant component plus isolated vertices
$p = \sqrt{\frac{2 \ln n}{n}}$	Diameter two
$p = \frac{\ln n}{n}$	Disappearance of isolated vertices Appearance of Hamilton circuit Diameter $O(\log n)$
$p = \frac{1}{2}$	Clique of size $(2 - \epsilon) \ln n$

Phase Transition

KAIST AI
Graduate School of AI

Definition

1. If there exists a function $p(n)$ such that
   when $\lim_{n\to\infty} \frac{p_1(n)}{p(n)} = 0$, $G(n, p_1(n))$ almost surely does not have the property, and
   when $\lim_{n\to\infty} \frac{p(n)}{p_2(n)} = 0$, $G(n, p_2(n))$ almost surely has the property,
   then we say that a phase transition occurs, and $p(n)$ is the threshold.
2. If for $cp(n), c < 1$, the graph almost surely does not have the property and for $cp(n), c > 1$, the graph almost surely has the property, then $p(n)$ is called a sharp threshold.

First moment and Second moment method

- Denote $x(n)$ the number of occurrences of an item in a random graph.
- By Markov's inequality, i.e., $Pr(x \geq a) \leq \frac{1}{a}E(x)$, which implies that $Pr(x(n) \geq 1) \leq E[x(n)]$.
- This is called the first moment method.
- When $E[x(n)]$ goes to infinity, if we have $Var(x(n)) = o(E[x(n)]^2)$, then $x(n)$ is almost surely greater than zero.
- This is called the second moment method.

First moment and Second moment method

KAIST AI
Graduate School of AI

Theorem. (Second Moment method)

Let $x(n)$ be a random variable with $E[x] > 0$. If

$$Var(x) = o(E^2(x))$$

then x is almost surely greater than zero.

Proof. If $E(x) > 0$, then for x to be less than or equal to zero, it must differ from its expected value by at least its expected value. Thus,

$$Prob(x \leq 0) \leq Prob\left(|x - E(x)| \geq E(x)\right)$$

By Chebyshev inequality,

$$Prob\left(|x - E(x)| \geq E(x)\right) \leq \frac{Var(x)}{E^2(x)} \to 0$$

Thus, $Prob(x \leq 0)$ goes to zero if $Var(x)$ is $o(E^2(x))$.

Roadmap

KAIST AI
Graduate School of AI

(1) The $G(n, p)$ Model

(2) Phase Transition

(3) Giant Component

(4) Cycles and Full Connectivity

(5) Phase Transition for Increasing Properties

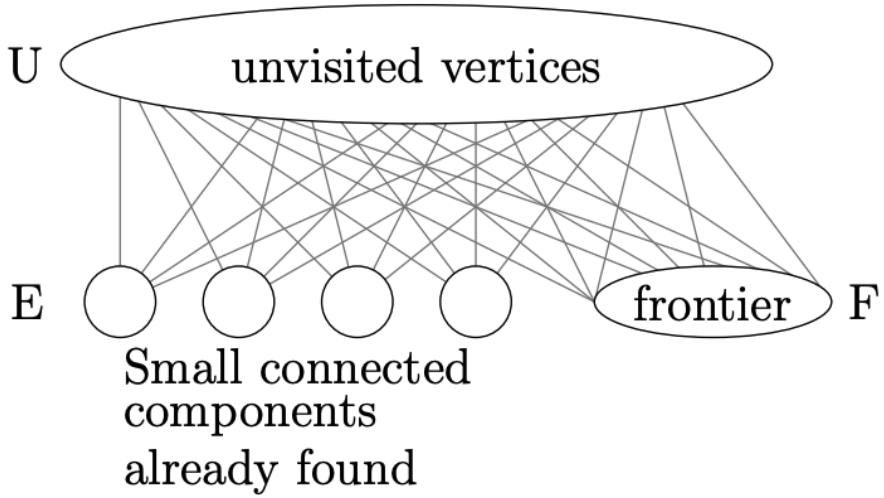
(6) Branching Processes

Giant Component

- Consider $G(n, p)$ for $p = \frac{1+\varepsilon}{n}$ where ε is a constant greater than zero.
- We now show that with high probability, such a graph contains a giant component, namely a component of size $\Omega(n)$.
- Moreover, with high probability, the graph contains only one such component, and all other components are much smaller, of size only $O(\log n)$.

Existence of a Giant Component

- To see that with high probability the graph has a giant component, do a depth first search (DFS) on $G(n,p)$ where $p = (1+\varepsilon)/n$ with $0 < \varepsilon < 1/8$.
- To perform the DFS, generate $\binom{n}{2}$ Bernoulli(p) independent random bits and answer the t^{th} edge query according to the t^{th} bit.
- As the DFS proceeds, let
  - E = set of fully explored vertices whose exploration is complete
  - U = set of unvisited vertices
  - F = frontier of visited and still being explored vertices



Existence of a Giant Component

- Intuitively, after $\varepsilon n^2/2$ edge queries a large number of edges must have been found since $p = \frac{1+\varepsilon}{n}$.
- None of theses can connect components already found with U, and we will use this to show that with high probability the frontier must be large.
- Since the frontier will be in a connected component, a giant component exists with high probability.

Lemma

After $\varepsilon n^2/2$ edge queries, with high probability $|E| < n/3$.

Proof. See page 262 of Textbook B.

Existence of a Giant Component

- The frontier vertices in the search of a connected component are all in the component being searched.
- Thus, if at any time the frontier set has $\Omega(n)$ vertices there exists a giant component.

Lemma

After $\varepsilon n^2/2$ edge queries, with high probability the set F consists of at least $\varepsilon^2 n/30$ vertices.

Proof. After $\varepsilon n^2/2$ queries, say $|F| < \varepsilon n^2/30$. Thus

$$|U| = n - |E| - |F| = n - \frac{n}{3} - \frac{\varepsilon^2 n}{30} \geq 1$$

The expected number of yes answers so far is $p\varepsilon n^2/2 = (1+\varepsilon)\varepsilon n/2$ and with high probability, the number of yes answers is at least $(\varepsilon n/2) + (\varepsilon^2 n/3)$.

Existence of a Giant Component

Proof cont'd So,

$$|E| + |F| \geq \frac{\varepsilon n}{2} + \frac{\varepsilon^2 n}{3} \Longrightarrow |E| > \frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}.$$

We must have $|E||U| \leq \varepsilon n^2/2$.
Now, $|E||U| = |E|(n - |E| - |F|)$ increases as $|E|$ increases from $\frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}$ to $n/3$, so we have

$$|E||U| \geq \left(\frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}\right)\left(n - \frac{\varepsilon n}{2} - \frac{3\varepsilon^2 n}{10} - \frac{\varepsilon^2 n}{30}\right) > \frac{\varepsilon n^2}{2},$$

a contradiction. □

Roadmap

KAIST AI
Graduate School of AI

(1) The $G(n, p)$ Model

(2) Phase Transition

(3) Giant Component

(4) Cycles and Full Connectivity

(5) Phase Transition for Increasing Properties

(6) Branching Processes

Cycles and Full Connectivity

KAIST AI
Graduate School of AI

- Consider when *cycles* form and when the graph becomes *fully connected*.
- For both of these problems, we look at each subset of k vertices and see when they form either a cycle or when they form a connected component.

Emergence of Cycles

- The emergence of cycles in $G(n,p)$ has a threshold when $p = 1/n$.

Theorem.

The threshold for the existence of cycles in $G(n,p)$ is $p = 1/n$.

Proof. Let x be the number of cycles in $G(n,p)$. To form a cycle of length k, the vertices can be selected in $\binom{n}{k}$ ways. Given the k vertices of the cycle, they can be ordered by arbitrarily selecting a first vertex, then a second vertex in one of $k-1$ ways, a third in one of $k-2$ ways, etc. Since a cycle and its reversal are the same cycle, divide by 2. Thus, there are $\binom{n}{k}\frac{(k-1)!}{2}$ possible cycles of length k and

$$E(x) = \sum_{k=3}^{n} \binom{n}{k} \frac{(k-1)!}{2} p^k \leq \sum_{k=3}^{n} \frac{n^k}{2k} p^k \leq \sum_{k=3}^{n} (np)^k = (np)^3 \frac{1-(np)^{n-2}}{1-np} \leq 2(np)^3$$

provided that $np < 1/2$. When p is asymptotically less than $1/n$, then $\lim_{n\to\infty} np = 0$ and $\lim_{n\to\infty} \sum_{k=3}^{n} (np)^k = 0$. So, as n goes to infinity, $E(x)$ goes to zero. Thus, the graph almost surely has no cycles by the first moment method. A second moment argument can be used to show that for $p = d/n$, $d > 1$, a graph will have a cycle with probability tending to one. □

Full Connectivity

- As p increases from $p = 0$, small components form.
- At $p = 1/n$ a giant component emerges and swallows up smaller components, starting with the larger components and ending up swallowing isolated vertices forming a single connected component at $p = \frac{\ln n}{n}$, at which point the graph becomes connected.

Lemma

The expected number of connected components of size k in $G(n, p)$ is at most

$$\binom{n}{k} k^{k-2} p^{k-1} (1-p)^{kn-k^2}$$

Proof. See page 267 of Textbook B.

Full Connectivity

- We now prove that for $p = c\frac{\ln n}{n}$ with $c > 1/2$, almost surely there are only isolated vertices and a giant component. For $c > 1$, almost surely the graph is connected.

Theorem.

The expected number of connected components of size k in $G(n, p)$ is at most

$$\binom{n}{k} k^{k-2} p^{k-1} (1-p)^{kn-k^2}$$

Proof. See page 267-268 of Textbook B.

Roadmap

KAIST AI
Graduate School of AI

(1) The $G(n, p)$ Model

(2) Phase Transition

(3) Giant Component

(4) Cycles and Full Connectivity

(5) Phase Transition for Increasing Properties

(6) Branching Processes

Phase Transition for Increasing Properties

KAIST AI
Graduate School of AI

- For many graph properties such as connectivity, having no isolated vertices, having a cycle, etc., the probability of a graph having the property increases as edges are added to the graph.
- Such a property is called an increasing property.
- Q is an *increasing property* of graphs if when a graph G has the property, any graph obtained by adding edges to G also have the property.
- Any increasing property has a threshold, although not necessarily a sharp one.

Phase Transition for Increasing Properties

KAIST AI
Graduate School of AI

- The notion of increasing property is defined in terms of *adding edges*.

Lemma

If Q is an increasing property of graphs and $0 \leq p \leq q \leq 1$, then the probability of $G(n, q)$ has property Q is greater than or equal to the probability that $G(n, p)$ has property Q.

Proof. Generate $G(n, q)$ as follows. First generate $G(n, p)$. Then, independently generate another graph $G(n, \frac{q-p}{1-p})$ and take the union by including an edge if either of the two graphs has the edge. Call the resulting graph H. The graph H has the same distribution as $G(n, q)$. This follows since the probability that an edge is in H is $p + (1-p)\frac{q-p}{1-p} = q$, and, clearly, the edges of H are independent. □

Remark. The lemma follows since whenever $G(n, p)$ has the property Q, H also has the property Q.

Phase Transition for Increasing Properties

KAIST AI
Graduate School of AI

- We now introduce a notion called replication.
- An m-fold replication of $G(n,p)$ is a random graph obtained as follows:
  1. Generate m independent copies of $G(n,p)$ on the same set of vertices
  2. Include an edge in the m-fold replication if the edge is in any one of the m copies
- The resulting random graph has the same distribution as $G(n,q)$ where $q = 1-(1-p)^m$, since the probability that a particular edge is not in the m-fold replication is the product of probabilities that is not in any of the m copies.
- If the m-fold replication of $G(n,p)$ does not have an increasing property then none of the m copies of $G(n,p)$ has the property. (The converse is not true)
- Since Q is an increasing property and $q = 1-(1-p)^m \leq 1-(1-mp) = mp$,

$$Prob(G(n,mp) \text{ has } Q) \geq Prob(G(n,q) \text{ has } Q) \tag{1}$$

Phase Transition for Increasing Properties

- Every increasing property Q has a phase transition.

Theorem.

Each increasing property Q of $G(n,p)$ has a phase transition at $p(n)$, where for each n, $p(n)$ is the minimum real number a_n for which the probability that $G(n,a_n)$ has property Q is $1/2$.

Proof. Let $p_0(n)$ be any function such that

$$\lim_{n\to\infty} \frac{p_0(n)}{p(n)} = 0.$$

We assert that almost surely $G(n,p_0)$ does not have the property Q. Suppose for contradiction, that this is not true, i.e. the probability that $G(n,p_0)$ has the property Q does not converge to zero. By the definition of a limit, there exists $\varepsilon > 0$ for which the probability that $G(n,p_0)$ has a property Q is at least ε on an infinite set I on n.

Phase Transition for Increasing Properties

Proof cont'd Let $m = \lceil (1/\varepsilon \rceil$. Let $G(n, q)$ be the m-fold replication of $G(n, p_0)$. The probability that $G(n, q)$ does not have Q is at most $(1 - \varepsilon)^m \leq e^{-1} \leq 1/2$ for all $n \in I$. For these n,

$$Prob(G(n, mp_0) \text{ has } Q) \geq Prob(G(n, q) \text{ has } Q) \geq 1/2.$$

Since $p(n)$ is the minimum real number a_n for which the probability that $G(n, a_n)$ has property Q is $1/2$, it must be that $mp_0(n) \geq p(n)$. This implies that $\frac{p_0(n)}{p(n)}$ is at least $1/m$ infinitely often, contradicting the hypothesis that $\lim_{n\to\infty} \frac{p_0(n)}{p(n)} = 0$.

A symmetric argument shows that for any $p_1(n)$ such that $\lim_{n\to\infty} \frac{p(n)}{p_1(n)} = 0$, $G(n, p_1)$ almost surely has property Q. □

Roadmap

KAIST AI
Graduate School of AI

(1) The $G(n, p)$ Model

(2) Phase Transition

(3) Giant Component

(4) Cycles and Full Connectivity

(5) Phase Transition for Increasing Properties

(6) Branching Processes

Branching Processes

- A branching process is a method for creating a random tree.
- Starting with the root node, each node has a probability distribution for the number of its children.
- Consider a simple case of a branching process where the distribution of the number of children at each node in the tree in the same.

Question: What is the probability that the tree is finite, i.e., the probability that the branching process dies out? This is called the extinction probability.

Branching Processes

- An important tool in analysis of branching process is the *generating function*.
- The generating function for a nonnegative integer valued random variable y is $f(x) = \sum_{i=0}^{\infty} p_i x^i$ where p_i is the probability that y equals i.
- Let the random variable z_j be the number of children in the j^{th} generation and let $f_j(x)$ be the generating function for z_j.
- Then $f_1(x) = f(x)$ is the generating function for the first generation where $f(x)$ is the generating function for the number of children at a node in the tree.
- The generating function for the 2^{th} generation is $f_2(x) = f(f(x))$.
- In general, the generating function for the $j + 1^{st}$ generation is $f_{j+1}(x) = f_j(f(x))$.

Branching Processes

- We can observe two things:
  1. The generating function for the sum of two identically distributed integer valued random variables x_1 and x_2 is the square of their generating function

$$f^2(x) = p_0^2 + (p_0p_1 + p_1p_0)x + (p_0p_2 + p_1p_1 + p_2p_0)x^2 + \cdots .$$

     In general, the generating function for the sum of i independent random variables, each with generating function $f(x)$, is $f^i(x)$.

  2. The coefficient of x^i in $f_j(x)$ is the probability of there being i children in the j^{th} generation. Thus, the generating function for the $j + 1^{st}$ generation, given i children in the j^{th} generation, is $f^i(x)$. The generating function for the $j + 1^{st}$ generation is given by

$$f_{j+1}(x) = \sum_{i=0}^{\infty} Prob(z_j = i) f^i(x)$$

     If f_{j+1} is obtained by substituting $f(x)$ for x in $f_j(x)$.

Branching Processes

- Let q be the probability that the branching process dies out.
- If there are i children in the first generation, then each of the i subtrees must die out and this occurs with probability q^i.
- Thus, q equals the summation over all values of i of the product of the probability of i children times the probability that i subtrees will die out.
- This gives $q = \sum_{i=0}^{\infty} p_i q^i$.
- Thus, q is the root of $x = \sum_{i=0}^{\infty} p_i x^i$, that is $x = f(x)$.
- This suggests focusing on roots of the equation $f(x) = x$ in the interval $[0, 1]$.

- Let $m = f'(1)$. Thus, m is the expected number of children of a node.

Lemma

Assume $m > 1$. Let q be the unique root of $f(x) = x$ in $[0, 1)$. In the limit as j goes to infinity, $f_j(x) = q$ for x in $[0, 1)$.

Proof. If $0 \leq x \leq q$, then $x < f(x) \leq f(q)$ and iterating this inequality

$$x < f_1(x) < f_2(x) < \cdots < f_j(x) < f(q) = q.$$

Clearly, the sequence converges and it must converge to a fixed point where $f(x) = x$. Similarly, if $q \leq x < 1$, then $f(q) \leq f(x) < x$ and iterating this inequality

$$x > f_1(x) > f_2(x) > \cdots > f_j(x) > f(q) = q.$$

In the limit as j goes to infinity $f_j(x) = q$ for all x, $0 \leq x < 1$. That is

$$\lim_{j \to \infty} f_j(x) = q + 0x + 0x^2 + \cdots$$

and there are no children with probability q and no finite number of children with probability zero. □

Branching Processes

KAIST AI
Graduate School of AI

Theorem.

Consider a tree generated by a branching process. Let $f(x)$ be the generating function for the number of children at each node.

1. If the expected number of children at each node is less than or equal to one, then the probability of extinction is one unless the probability of exactly one child is one.
2. If the expected number of children of each node is greater than one, then the probability of extinction is the unique solution to $f(x) = x$ in $[0, 1)$.

Proof. Let p_i be the probabability of i children at each node. Then,
$f(x) = p_0 + p_1 x + p_2 x^2 + \cdots$ is generating function for the number of children at each node and $f'(1) = p_1 + 2p_2 + 3p_3 + \cdots$ is the slope of $f(x)$ at $x = 1$.
Observe that $f'(1)$ is the expected number of children at each node.

Proof cont'd.

Since the expected number of children at each node is the slope of $f(x)$ at $x = 1$,
if the expected number of children is less than or equal to one,
the slope of $f(x)$ at $x = 1$ is less than or equal to one and the unique root of $f(x) = x$ in $(0, 1]$ is at $x = 1$ and
the probability of extinction is one unless $f'(1) = 1$ and $p_1 = 1$.
If $f'(1) = 1$ and $p_1 = 1$, $f(x) = x$ and the tree is an infinite degree one chain.
If the slope of $f(x)$ at $x = 1$ is greater than one, then the probability of extinction is the unique solution to $f(x) = x$ in $[0, 1)$. □

Expected Size of Extinct Families

- Consider that the expected size of an extinct family is finite, provided that $m \neq 1$.
- The expected size at extinction could conceivably infinite, if the probability of dying out did not decay fast enough.
- In such a case the expected size at extinction would be infinite even though the process dies out with probability one.
- The following lemma shows this does not happen.

Lemma

If the slope $m = f'(1)$ does not equal one, then the expected size of an extinct family is finite. If the slope m equals one and $p_1 = 1$, then the tree is an infinite degree one chain and there are no extinct families. If $m = 1$ and $p_1 < 1$, then the expected size of the extinct family is infinite.

Proof. See page 276-277 of Textbook B.

Questions?