Models of information propagation in online social networks: Epidemic processes and random graphs

Laurent Massoulié

Inria

January 20, 2021
Viral propagation of information and "information cascades"
Propagation on underlying graph (e.g. facebook’s "friendship graph", or Twitter’s "follower-followee" directed graph)

→ Epidemic models to understand viral propagation (and guide viral marketing strategies)
Assigns to each oriented edge \((i,j)\) a probability \(p_{ij}\).

\(i\) infected in slot \(t\) \(\Rightarrow\) infects each neighbor \(j\) with probability \(p_{ij}\) in slot \(t+1\) independently of everything else and is then Removed.

Questions of interest: Number of eventually infected nodes? As a function of set initially infected? Optimal choice of initial set of given size?
SIR epidemics: the Reed-Frost model

- Special case: complete graph on \( i \in [n] \) and homogeneous infection probabilities \( p_{ij} \equiv p \)
SIR epidemics: the Reed-Frost model

- Special case: complete graph on $i \in [n]$ and homogeneous infection probabilities $p_{ij} \equiv p$
- Associated model: Erdős-Rényi random graph $G(n, p)$: undirected graph on node set $[n]$. Edge $(i, j)$ present iff $\xi_{ij} = 1$ where $\{\xi_{ij}\}_{i<j}$: i.i.d., Bernoulli ($p$)
SIR epidemics: the Reed-Frost model

- Special case: complete graph on $i \in [n]$ and homogeneous infection probabilities $p_{ij} \equiv p$
- Associated model: Erdős-Rényi random graph $G(n, p)$: undirected graph on node set $[n]$. Edge $(i, j)$ present iff $\xi_{ij} = 1$ where $\{\xi_{ij}\}_{i < j}$: i.i.d., Bernoulli ($p$)
- From random graph to epidemic process: use $\xi_{ij}$ to determine if when the first of $i$ and $j$ gets infected, it infects the other
SIR epidemics: the Reed-Frost model

- Special case: complete graph on \( i \in [n] \) and homogeneous infection probabilities \( p_{ij} \equiv p \)
- Associated model: Erdős-Rényi random graph \( G(n, p) \): undirected graph on node set \([n]\). Edge \((i, j)\) present iff \( \xi_{ij} = 1 \) where \( \{\xi_{ij}\}_{i<j} \): i.i.d., Bernoulli \( (p) \)
- From random graph to epidemic process: use \( \xi_{ij} \) to determine if when the first of \( i \) and \( j \) gets infected, it infects the other

\[ \Rightarrow \text{For initial set } X_0 \text{ of infective nodes at time } 0, \text{ } i \text{ infected at time } t \text{ iff } d_G(X_0, i) = t \]

Set of nodes eventually infected: \( \bigcup_{i \in X_0} \Gamma(i) \) where \( \Gamma(i) \): graph's connected component including \( i \)
Seminal results by Erdős and Rényi (1959-1960)

- First phase transition: emergence of giant component
  - Tools: branching processes & Chernoff’s inequality

- Second phase transition: emergence of connectivity
  - Tools: 1st and 2nd moment methods; Poisson approximation
Towards Susceptible-Infected-Removed (SIR) epidemics: Galton-Watson branching process (1873)

Offspring distribution \( \{p_k\}_{k \in \mathbb{N}} \)

\( Z_k \) number of individuals per generation:

\[
Z_0 = 1, \quad Z_k = \sum_{m=1}^{Z_{k-1}} X_{m,k} \quad \text{where} \quad \{X_{m,k}\}_{m,k \geq 0}: \text{i.i.d.,} \sim \{p_k\}_{k \in \mathbb{N}}
\]

Quantities of interest: probability of extinction; in case of extinction, total population size
**Theorem**

**Extinction probability** $p_{\text{ext}}$: smallest root in $[0, 1]$ of $z = \phi(z)$ where 

$$\phi(z) = \mathbb{E}(z^X) = \sum_{k \geq 0} p_k z^k$$

- If $\mu := \mathbb{E}(X) < 1$ then $p_{\text{ext}} = 1$
- If $\mu = 1$ and $p_0 > 0$ then $p_{\text{ext}} = 1$
- If $\mu > 1$ then $p_{\text{ext}} < 1$

**Laurent Massoulié (Inria)**

Models of information propagation in online social networks: Epidemic processes and random graphs

January 20, 2021 7 / 24
Theorem

Extinction probability $p_{\text{ext}}$: smallest root in $[0, 1]$ of $z = \phi(z)$ where

$\phi(z) = E(z^X) = \sum_{k \geq 0} p_k z^k$

If $\mu := E(X) < 1$ then $p_{\text{ext}} = 1$

If $\mu = 1$ and $p_0 > 0$ then $p_{\text{ext}} = 1$

If $\mu > 1$ then $p_{\text{ext}} < 1$

Proof: $\{Z_k = 0\} \uparrow \{\text{Extinction}\}$; $P(Z_k = 0) = \phi_k(0)$ where

$\phi_k(z) = E(z^{Z_k})$

By induction $\phi_k(z) = \phi \circ \phi_{k-1}(z)$ hence $P(Z_k = 0) = \phi(P(Z_{k-1} = 0))$

$\Rightarrow$ by monotonicity of $\phi$ and $P(Z_0 = 0) = 0$, sequence increases to (necessarily smallest) fixed point.
Theorem

Extinction probability $p_{\text{ext}}$: smallest root in $[0, 1]$ of $z = \phi(z)$ where

\[
\phi(z) = \mathbb{E}(z^X) = \sum_{k \geq 0} p_k z^k
\]

If $\mu := \mathbb{E}(X) < 1$ then $p_{\text{ext}} = 1$

If $\mu = 1$ and $p_0 > 0$ then $p_{\text{ext}} = 1$

If $\mu > 1$ then $p_{\text{ext}} < 1$

Proof: $\{Z_k = 0\} \uparrow \{\text{Extinction}\}; \mathbb{P}(Z_k = 0) = \phi_k(0)$ where

\[
\phi_k(z) = \mathbb{E}(z^{Z_k})
\]

By induction $\phi_k(z) = \phi \circ \phi_{k-1}(z)$ hence $\mathbb{P}(Z_k = 0) = \phi(\mathbb{P}(Z_{k-1} = 0))$.

$\Rightarrow$ by monotonicity of $\phi$ and $\mathbb{P}(Z_0 = 0) = 0$, sequence increases to (necessarily smallest) fixed point.

$\mu$: slope of $\phi$ at $1^-$. By convexity of $\phi$, only fixed point: 1 if $\mu < 1$

By continuity of $\phi$, $\exists$ fixed point $< 1$ if $\mu > 1$

For $\mu = 1$, if $p_0 > 0$ then $\phi$ strictly convex hence only fixed point: 1; if $p_0 = 0$ then $p_{\text{ext}} = 0$. 

Fundamental example of phase transition

Special case $X \sim \text{Poisson}(\mu)$: $p_{\text{ext}} = e^{-\mu}(1 - p_{\text{ext}})$
Theorem

**Extinction probability** $p_{\text{ext}}$: smallest root in $[0, 1]$ of $z = \phi(z)$ where

$$\phi(z) = \mathbb{E}(z^X) = \sum_{k \geq 0} p_k z^k$$

If $\mu := \mathbb{E}(X) < 1$ then $p_{\text{ext}} = 1$
If $\mu = 1$ and $p_0 > 0$ then $p_{\text{ext}} = 1$
If $\mu > 1$ then $p_{\text{ext}} < 1$

Proof: $\{Z_k = 0\} \uparrow \{\text{Extinction}\}$; $\mathbb{P}(Z_k = 0) = \phi_k(0)$ where

$$\phi_k(z) = \mathbb{E}(z^{Z_k})$$

By induction $\phi_k(z) = \phi \circ \phi_{k-1}(z)$ hence $\mathbb{P}(Z_k = 0) = \phi(\mathbb{P}(Z_{k-1} = 0))$

$\Rightarrow$ by monotonicity of $\phi$ and $\mathbb{P}(Z_0 = 0) = 0$, sequence increases to (necessarily smallest) fixed point.

$\mu$: slope of $\phi$ at $1^-$. By convexity of $\phi$, only fixed point: 1 if $\mu < 1$

By continuity of $\phi$, $\exists$ fixed point $< 1$ if $\mu > 1$

For $\mu = 1$, if $p_0 > 0$ then $\phi$ strictly convex hence only fixed point: 1; if $p_0 = 0$ then $p_{\text{ext}} = 0$

Fundamental example of **phase transition**

Special case $X \sim \text{Poisson}(\mu)$: $p_{\text{ext}} = e^{-\mu(1-p_{\text{ext}})}$
Random walk exploration of Galton-Watson tree

Sequentially pick *active* node (whose children have not yet been sampled)
De-activate it and add its children to active set
Stop when active set empty (tree exploration complete)

Dynamics of $A_t$, number of active nodes at step $t$:

$$A_t = A_{t-1} - 1 + X_t$$

where $X_t$ independent of past exploration $\{A_s, X_s, s < t\}$ and distributed according to $\{p_k\}_{k \geq 0}$

Time $T$ at which exploration stops, i.e. $A_T = 0$ gives size of tree.

Indeed $A_t = 1 - t + X_1 + \ldots + X_t$ and $A_T = 0$ yield $T = 1 + X_1 + \ldots + X_T$.

Random walk can be pursued after time $T$ ⇒ Bound on population size: for continued RW $\{A_t\}_{t \geq 0}$,

$$P(T > t) = P(A_1, \ldots, A_t > 0) \leq P(A_t > 0) = P(\sum_{s=1}^{t} (X_s - 1) \geq 0)$$
Random walk exploration of Galton-Watson tree

Sequentially pick active node (whose children have not yet been sampled) 
De-activate it and add its children to active set 
Stop when active set empty (tree exploration complete)

- Dynamics of $A_t$, number of active nodes at step $t$:
  Random walk $A_t = A_{t-1} - 1 + X_t$ where $X_t$ independent of past 
  exploration $\{A_s, X_s, s < t\}$ and distributed according to $\{p_k\}_{k \geq 0}$
- Time $T$ at which exploration stops, i.e. $A_T = 0$ gives size of tree. 
  Indeed $A_t = 1 - t + X_1 + \ldots + X_t$ and $A_T = 0$ yield 
  $T = 1 + X_1 + \ldots + X_T$.
- Random walk can be pursued after time $T$
Random walk exploration of Galton-Watson tree

Sequentially pick *active* node (whose children have not yet been sampled)
De-activate it and add its children to active set
Stop when active set empty (tree exploration complete)

- Dynamics of $A_t$, number of active nodes at step $t$:
  Random walk $A_t = A_{t-1} - 1 + X_t$ where $X_t$ independent of past
  exploration $\{A_s, X_s, s < t\}$ and distributed according to $\{p_k\}_{k \geq 0}$

- Time $T$ at which exploration stops, i.e. $A_T = 0$ gives size of tree.
  Indeed $A_t = 1 - t + X_1 + \ldots + X_t$ and $A_T = 0$ yield
  $T = 1 + X_1 + \ldots + X_T$.

- Random walk can be pursued after time $T$

  $\Rightarrow$ Bound on population size: for continued RW $\{A_t\}_{t \geq 0}$,
  $\mathbb{P}(T > t) = \mathbb{P}(A_1, \ldots, A_t > 0) \leq \mathbb{P}(A_t > 0) = \mathbb{P}(\sum_{s=1}^t (X_s - 1) \geq 0)$
Control of fluctuations: Chernoff’s inequality

- **Markov’s inequality**: random variable $X \geq 0$, $a > 0 \Rightarrow \mathbb{P}(X \geq a) \leq \mathbb{E}(X)/a$
Control of fluctuations: Chernoff’s inequality

- **Markov’s inequality**: random variable $X \geq 0$, $a > 0 \Rightarrow \mathbb{P}(X \geq a) \leq \mathbb{E}(X)/a$

- **Bienaymé-Tchebichev’s inequality**: random variable $X \in \mathbb{R}$: $\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \text{Var}(X)/a^2$
Control of fluctuations: Chernoff’s inequality

- **Markov’s inequality**: random variable $X \geq 0$, $a > 0 \Rightarrow P(X \geq a) \leq \mathbb{E}(X)/a$

- **Bienaymé-Tchebitchev’s inequality**: random variable $X \in \mathbb{R}$: $P(|X - \mathbb{E}(X)| \geq a) \leq \text{Var}(X)/a^2$

- **Exponential version**: for $\theta > 0$, $P(X \geq t) \leq \mathbb{E}(e^{\theta X})e^{-\theta t}$ i.e. finite exponential moments yield exponentially decaying control of tail probabilities
Chernoff’s inequality and bounds on population size

Theorem
For i.i.d. $X_s$, $\mathbb{P}(\sum_{s=1}^{t} X_s \geq at) \leq e^{-th(a)}$ where $h(a) := \sup_{\theta > 0}[\theta a - \ln(\mathbb{E}(e^{\theta X_1}))]$
Chernoff’s inequality and bounds on population size

**Theorem**

For i.i.d. $X_s$, $\mathbb{P}(\sum_{s=1}^{t} X_s \geq at) \leq e^{-th(a)}$ where 
$h(a) := \sup_{\theta > 0} [\theta a - \ln(\mathbb{E}(e^{\theta X_1}))]$

Non-trivial exponential bound when $a > \mathbb{E}(X_1)$ and $\exists \epsilon > 0 : \mathbb{E}e^{\epsilon X_1} < +\infty$
Chernoff’s inequality and bounds on population size

**Theorem**

For i.i.d. $X_s$, $\Pr(\sum_{s=1}^{t} X_s \geq at) \leq e^{-th(a)}$ where $h(a) := \sup_{\theta > 0} [\theta a - \ln(\mathbb{E}(e^{\theta X_1}))]$.

Non-trivial exponential bound when $a > \mathbb{E}(X_1)$ and $\exists \epsilon > 0 : \mathbb{E}e^{\epsilon X_1} < +\infty$.

Application to Galton-Watson process:
$\Pr(T > t) \leq e^{-th(1)}$ exponentially decaying if $\mathbb{E}(X_1) < 1$ and $X_1$ admits finite exponential moments.
Chernoff’s inequality and bounds on population size

**Theorem**

For i.i.d. $X_s$, $\mathbb{P}(\sum_{s=1}^{t} X_s \geq a t) \leq e^{-th(a)}$ where $h(a) := \sup_{\theta > 0} [\theta a - \ln(\mathbb{E}(e^{\theta X_1}))]$

Non-trivial exponential bound when $a > \mathbb{E}(X_1)$ and $\exists \epsilon > 0 : \mathbb{E}e^{\epsilon X_1} < +\infty$

Application to Galton-Watson process:
$\mathbb{P}(T > t) \leq e^{-th(1)}$ exponentially decaying if $\mathbb{E}(X_1) < 1$ and $X_1$ admits finite exponential moments.

Case of Poisson random variables, parameter $\mu > 0$, $a > \mu$:
$h_\mu(a) = \sup_{\theta > 0} [\theta a - \mu(e^{\theta} - 1)]$
Gives $\theta = \ln(a/\mu)$, $h_\mu(a) = \mu h_1(a/\mu)$
with $h_1(x) = x \ln(x) - x + 1$
Emergence of giant component

Analysis of graph’s connected components: let $C(i)$: size of $i$-th largest
connected component (in number of nodes) in $G(n, p)$

Theorem

Let $p = \lambda/n$ for fixed $\lambda > 0$

**Sub-critical case** ($\lambda < 1$): there exists $f(\lambda)$ such that

$$
\lim_{n \to \infty} \mathbb{P}(C(1) \leq f(\lambda) \ln(n)) = 1
$$

**Super-critical case** ($\lambda > 1$): there exists $g(\lambda)$ such that for all $\delta > 0$,

$$
\lim_{n \to \infty} \mathbb{P}\left(|\frac{C(1)}{n} - (1 - p_{\text{ext}})| \leq \delta, \ C(2) \leq g(\lambda) \ln(n)\right) = 1,
$$

where $p_{\text{ext}}$: extinction probability of Poisson ($\lambda$) branching process, i.e. smallest root of $x = e^{\lambda(x-1)}$ in $[0, 1]$
Interpretation

**Sub-critical regime**: Only logarithmically sized components i.e. no global outbreak

**Super-critical regime**: with probability $1 - p_{\text{ext}}$, epidemics started from randomly selected node reaches $n[1 - p_{\text{ext}} + o(1)]$ others, i.e. macroscopic outbreak

Note: only one giant component, others still logarithmic
Sub-critical regime

- Exploration of connected component $\Gamma(i_0)$: initialized with active set $A_0 = \{i_0\}$ and killed set $B_0 = \emptyset$

- At time $t$ pick $j_t \in A_{t-1}$, kill it and activate its neighbours not yet activated (set $D_t$)
  \[ A_t = A_{t-1} \setminus \{j_t\} \cup D_t, \quad B_t = B_{t-1} \cup \{j_t\} \]
Sub-critical regime

- Exploration of connected component $\Gamma(i_0)$: initialized with active set $A_0 = \{i_0\}$ and killed set $B_0 = \emptyset$

- At time $t$ pick $j_t \in A_{t-1}$, kill it and activate its neighbours not yet activated (set $D_t$)
  $\Rightarrow A_t = A_{t-1} \setminus \{j_t\} \cup D_t, B_t = B_{t-1} \cup \{j_t\}$

- Notation: $A_t = |A_t|, D_t = |D_t| \Rightarrow A_t = 1 - t + D_1 + \cdots + D_t$
Sub-critical regime

- Exploration of connected component $\Gamma(i_0)$: initialized with active set $A_0 = \{i_0\}$ and killed set $B_0 = \emptyset$
- At time $t$ pick $j_t \in A_{t-1}$, kill it and activate its neighbours not yet activated (set $D_t$)
  \[ A_t = A_{t-1} \setminus \{j_t\} \cup D_t, \quad B_t = B_{t-1} \cup \{j_t\} \]
- Notation: $A_t = |A_t|$, $D_t = |D_t| \Rightarrow A_t = 1 - t + D_1 + \cdots + D_t$
- Conditionally on $F_{t-1} = \sigma(A_1, \ldots, A_{t-1})$,
  \[ D_t \sim \text{Bin}(p, n - 1 - D_1 - \cdots - D_{t-1}) \]
Sub-critical regime

- Exploration of connected component $\Gamma(i_0)$: initialized with active set $A_0 = \{i_0\}$ and killed set $B_0 = \emptyset$

- At time $t$ pick $j_t \in A_{t-1}$, kill it and activate its neighbours not yet activated (set $D_t$)
  \[ A_t = A_{t-1} \setminus \{j_t\} \cup D_t, \quad B_t = B_{t-1} \cup \{j_t\} \]

- Notation: $A_t = |A_t|, \quad D_t = |D_t| \Rightarrow A_t = 1 - t + D_1 + \cdots + D_t$

- Conditionally on $\mathcal{F}_{t-1} = \sigma(A_1, \ldots, A_{t-1})$, $D_t \sim \text{Bin}(p, \, n - 1 - D_1 - \cdots - D_{t-1})$

- Size $C$ of connected component:
  \[ C = \inf\{t > 0 : A_t = 0\} \]
Sub-critical regime, continued

- Processes $\{A_t\}, \{D_t\}$ can be extended after end of component’s exploration

- Upper bound:

$$\mathbb{P}(C > k) = \mathbb{P}(A_1, \ldots, A_k > 0) \leq \mathbb{P}(A_k > 0)$$
Sub-critical regime, continued

- Processes \( \{A_t\}, \{D_t\} \) can be extended after end of component’s exploration

- Upper bound:
  \[
  \mathbb{P}(C > k) = \mathbb{P}(A_1, \ldots, A_k > 0) \leq \mathbb{P}(A_k > 0)
  \]

- Chernoff’s bounding technique: \( \mathbb{P}(A_k > 0) \leq e^{-kh(1)} \)
  
  where \( h(x) = \lambda h_1(x/\lambda), \quad h_1(x) = x \ln(x) - x + 1 \): Chernoff’s exponent for Poisson (\( \lambda \)) random variable
Sub-critical regime, continued

- Processes $\{A_t\}, \{D_t\}$ can be extended after end of component’s exploration.

- Upper bound:

$$\mathbb{P}(C > k) = \mathbb{P}(A_1, \ldots, A_k > 0) \leq \mathbb{P}(A_k > 0)$$

- Chernoff’s bounding technique: $\mathbb{P}(A_k > 0) \leq e^{-kh(1)}$

  where $h(x) = \lambda h_1(x/\lambda), \ h_1(x) = x \ln(x) - x + 1$: Chernoff’s exponent for Poisson ($\lambda$) random variable.

- Union bound allows to conclude.
Super-critical regime $\lambda > 1$

**Lemma**

For any $k > 0$, $d_1, \ldots, d_k \in \mathbb{N}^k$, $\lim_{n \to \infty} \mathbb{P}(D_1^k = d_1^k) = \prod_{s=1}^{k} e^{-\lambda} \frac{\lambda^{d_s}}{d_s!}$, hence $\lim_{n \to \infty} \mathbb{P}(C \leq k) = \mathbb{P}(Z \leq k) \leq p_{ext}$

where $Z$: total population of Poisson $(\lambda)$ branching process

Additional technical steps involved to characterize sizes of connected components in super-critical regime, see notes.
Connectivity

By previous result: for fixed $\lambda > 1$, giant component of size $\sim n(1 - p_{ext})$

For fixed $\lambda$, graph disconnected $\Rightarrow$ Under what regime is graph connected?
Connectivity

By previous result: for fixed $\lambda > 1$, giant component of size $\sim n(1 - p_{\text{ext}})$
For fixed $\lambda$, graph disconnected $\Rightarrow$ Under what regime is graph connected?

**Theorem**

*For fixed $c \in \mathbb{R}$, assume $np = \ln(n) + c$. Then $\lim_{n \to \infty} \mathbb{P}(G(n, p) \text{ connected}) = e^{-e^{-c}}$*
Connectivity

By previous result: for fixed $\lambda > 1$, giant component of size $\sim n(1 - p_{ext})$
For fixed $\lambda$, graph disconnected $\Rightarrow$ Under what regime is graph connected?

**Theorem**

For fixed $c \in \mathbb{R}$, assume $np = \ln(n) + c$.
Then $\lim_{n \to \infty} \mathbb{P}(G(n, p) \text{ connected}) = e^{-e^{-c}}$

**Corollary**

If $np - \ln(n) \to +\infty$, then $\lim_{n \to \infty} \mathbb{P}(G(n, p) \text{ connected}) = 1$
If $np - \ln(n) \to -\infty$, then $\lim_{n \to \infty} \mathbb{P}(G(n, p) \text{ connected}) = 0$
Proof strategy

- Show that number of isolated nodes (i.e. nodes of degree 0) admits asymptotically Poisson ($e^{-c}$) distribution [Poisson approximation method],

  hence $\lim_{n \to \infty} P(A) = e^{-e^{-c}}$ where $A = \{\text{no isolated vertices in } G(n, p)\}$
Proof strategy

- Show that number of isolated nodes (i.e. nodes of degree 0) admits asymptotically Poisson ($e^{-c}$) distribution [Poisson approximation method],

  hence $\lim_{n \to \infty} \mathbb{P}(A) = e^{-e^{-c}}$ where $A = \{\text{no isolated vertices in } G(n, p)\}$

- Show that $\lim_{n \to \infty} \mathbb{P}(B) = 0$ where $B = \{\exists \text{ connected component of size } k \in \{2, \ldots, n/2\}\}$
Proof strategy

- Show that number of **isolated nodes** (i.e. nodes of degree 0) admits asymptotically Poisson \((e^{-c})\) distribution [Poisson approximation method],

  hence \(\lim_{n \to \infty} P(A) = e^{-e^{-c}}\) where
  \[A = \{\text{no isolated vertices in } G(n, p)\}\]

- Show that \(\lim_{n \to \infty} P(B) = 0\) where
  \[B = \{\exists \text{ connected component of size } k \in \{2, \ldots, n/2\}\}\]

- Use bounds

  \[P(A) - P(B) \leq P(G(n, p) \text{ connected}) = P(A \cap \overline{B}) \leq P(A)\]
Basic tools: the first and second moment methods

Let $Z_u, u \in V$ be indicators of events and $X = \sum_{u \in V} Z_u$.

First moment method:

$$P(\exists u \in V: Z_u = 1) \leq \sum_{u \in V} E(Z_u) = E(X),$$

hence "with high probability" none of these events occurs if

$$\lim_{n \to \infty} E(X) = 0.$$ 

Application: with high probability no isolated node in $G(n, p)$ if

$$\lim_{n \to \infty} \left[ np - \ln(n) \right] = +\infty.$$ 

Second moment method:

$$P(\forall u \in V, Z_u = 0) = P(X = 0) \leq \text{Var}(X) \frac{E(X)}{E(X)^2},$$

Hence if $\text{Var}(X) = o(E(X)^2)$, then with high probability some event occurs.

Application: with high probability there is some isolated node in $G(n, p)$ if

$$\lim_{n \to \infty} \left[ np - \ln(n) \right] = -\infty.$$
Basic tools: the first and second moment methods

Let $Z_u, u \in V$ be indicators of events and $X = \sum_{u \in V} Z_u$.

**First moment method:** $\mathbb{P}(\exists u \in V : Z_u = 1) \leq \sum_{u \in V} \mathbb{E}(Z_u) = \mathbb{E}(X)$, hence “with high probability” none of these events occurs if $\lim_{n \to \infty} \mathbb{E}(X) = 0$.

**Second moment method:** $\mathbb{P}(\forall u \in V, Z_u = 0) = \mathbb{P}(X = 0) \leq \text{Var}(X) \mathbb{E}(X)$. Hence if $\text{Var}(X) = o(\mathbb{E}(X)^2)$, then with high probability some event occurs.

Application: with high probability no isolated node in $G(n, p)$ if $\lim_{n \to \infty} [np - \ln(n)] = +\infty$.

Application: with high probability there is some isolated node in $G(n, p)$ if $\lim_{n \to \infty} [np - \ln(n)] = -\infty$.
Basic tools: the first and second moment methods

Let $Z_u, u \in V$ be indicators of events and $X = \sum_{u \in V} Z_u$.

**First moment method:** $\mathbb{P}(\exists u \in V : Z_u = 1) \leq \sum_{u \in V} \mathbb{E}(Z_u) = \mathbb{E}(X)$, hence “with high probability” none of these events occurs if $\lim_{n \to \infty} \mathbb{E}(X) = 0$.

Application: with high probability no isolated node in $G(n, p)$ if $\lim_{n \to \infty} [np - \ln(n)] = +\infty$.

**Second moment method:** $\mathbb{P}(\forall u \in V, Z_u = 0) = \mathbb{P}(X = 0) \leq \text{Var}(X) \mathbb{E}(X)^2$.

Hence if $\text{Var}(X) = o(\mathbb{E}(X)^2)$, then with high probability some event occurs.

Application: with high probability there is some isolated node in $G(n, p)$ if $\lim_{n \to \infty} [np - \ln(n)] = -\infty$. 
Basic tools: the first and second moment methods

Let $Z_u, u \in V$ be indicators of events and $X = \sum_{u \in V} Z_u$.

**First moment method**: $\mathbb{P}(\exists u \in V : Z_u = 1) \leq \sum_{u \in V} \mathbb{E}(Z_u) = \mathbb{E}(X)$, hence “with high probability” none of these events occurs if $\lim_{n \to \infty} \mathbb{E}(X) = 0$.

Application: with high probability no isolated node in $G(n, p)$ if $\lim_{n \to \infty} [np - \ln(n)] = +\infty$.

**Second moment method**: $\mathbb{P}(\forall u \in V, Z_u = 0) = \mathbb{P}(X = 0) \leq \frac{\text{Var}(X)}{\mathbb{E}(X)^2}$.

Hence if $\text{Var}(X) = o(\mathbb{E}(X)^2)$, then with high probability some event occurs.
Basic tools: the first and second moment methods

Let $Z_u, u \in V$ be indicators of events and $X = \sum_{u \in V} Z_u$.

First moment method: $\mathbb{P}(\exists u \in V : Z_u = 1) \leq \sum_{u \in V} \mathbb{E}(Z_u) = \mathbb{E}(X)$, hence “with high probability” none of these events occurs if $\lim_{n \to \infty} \mathbb{E}(X) = 0$.

Application: with high probability no isolated node in $G(n, p)$ if $\lim_{n \to \infty} [np - \ln(n)] = +\infty$.

Second moment method: $\mathbb{P}(\forall u \in V, Z_u = 0) = \mathbb{P}(X = 0) \leq \frac{\text{Var}(X)}{\mathbb{E}(X)^2}$.

Hence if $\text{Var}(X) = o(\mathbb{E}(X)^2)$, then with high probability some event occurs.

Application: with high probability there is some isolated node in $G(n, p)$ if $\lim_{n \to \infty} [np - \ln(n)] = -\infty$. 
Variation distance

Definition

Variation distance between two probability measures $\mu, \nu$ on $(\Omega, \mathcal{F})$: $d_{\text{var}}(\mu, \nu) = 2 \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|$
Variation distance

**Definition**

Variation distance between two probability measures \( \mu, \nu \) on \( (\Omega, \mathcal{F}) \):

\[
d_{\text{var}}(\mu, \nu) = 2 \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|
\]

Alternative characterization: if \( \mu, \nu \) admit densities \( \frac{d\mu}{d\pi}, \frac{d\nu}{d\pi} \) with respect to measure \( \pi \) (e.g., \( \pi = \mu + \nu \)) then

\[
d_{\text{var}}(\mu, \nu) = \int_{\Omega} \left| \frac{d\mu}{d\pi} - \frac{d\nu}{d\pi} \right| d\pi
\]

In particular for \( \Omega = \mathbb{N} \) and \( \pi = \sum_{n \in \mathbb{N}} \delta_n \), \( d_{\text{var}}(\mu, \nu) = \sum_{n \in \mathbb{N}} |\mu_n - \nu_n| \)
Variation distance

**Definition**

Variation distance between two probability measures $\mu, \nu$ on $(\Omega, \mathcal{F})$:

$$d_{\text{var}}(\mu, \nu) = 2 \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|$$

Alternative characterization: if $\mu, \nu$ admit densities $\frac{d\mu}{d\pi}, \frac{d\nu}{d\pi}$ with respect to measure $\pi$ (e.g., $\pi = \mu + \nu$) then

$$d_{\text{var}}(\mu, \nu) = \int_{\Omega} \left| \frac{d\mu}{d\pi} - \frac{d\nu}{d\pi} \right| d\pi$$

In particular for $\Omega = \mathbb{N}$ and $\pi = \sum_{n \in \mathbb{N}} \delta_n$, $d_{\text{var}}(\mu, \nu) = \sum_{n \in \mathbb{N}} |\mu_n - \nu_n|$

**Definition**

$\{\mu^{(n)}\}_{n \in \mathbb{N}}$ converges in variation to $\mu$ iff $\lim_{n \to \infty} d_{\text{var}}(\mu^{(n)}, \mu) = 0$

A strong form of convergence (implies convergence in distribution)
Poisson approximation: the Stein-Chen method

**Theorem**

Let $Z_u \in \{0, 1\}$, $u \in V$, $X = \sum_{u \in V} Z_u$.

Denote $\pi_u = \mathbb{E}(Z_u)$, $\lambda = \mathbb{E}(X) = \sum_{u \in V} \pi_u$. 

Assume $\exists \{Z_{uv}\}_{u,v \in V, v \neq u}$ such that $\forall u \in V$, $P(\{Z_{uv}\}_{v \neq u} \in \cdot | Z_u = 1)$. Then:

$$d_{var}(X, \text{Poisson}(\lambda)) \leq 2 \min(1, 1/\lambda) \sum_{u \in V} \pi_u \left[ \pi_u + \sum_{v \neq u} \mathbb{E}|Z_{uv} - Z_v| \right].$$
Poisson approximation: the Stein-Chen method

**Theorem**

Let $Z_u \in \{0, 1\}$, $u \in V$, $X = \sum_{u \in V} Z_u$. Denote $\pi_u = \mathbb{E}(Z_u)$, $\lambda = \mathbb{E}(X) = \sum_{u \in V} \pi_u$. Assume $\exists \{Z_{uv}\}_{u,v \in V, v \neq u}$ such that

$$\forall u \in V, \mathbb{P}(\{Z_{uv}\}_{v \neq u} \in \cdot) = \mathbb{P}(\{Z_v\}_{v \neq u} \in \cdot | Z_u = 1).$$
Poisson approximation: the Stein-Chen method

**Theorem**

Let $Z_u \in \{0, 1\}$, $u \in V$, $X = \sum_{u \in V} Z_u$.

Denote $\pi_u = \mathbb{E}(Z_u)$, $\lambda = \mathbb{E}(X) = \sum_{u \in V} \pi_u$.

Assume $\exists \{Z_{uv}\}_{u,v \in V, v \neq u}$ such that

$$\forall u \in V, \mathbb{P}(\{Z_{uv}\}_{v \neq u} \in \cdot) = \mathbb{P}(\{Z_v\}_{v \neq u} \in \cdot | Z_u = 1).$$

Then:

$$d_{var}(X, \text{Poisson}(\lambda)) \leq 2 \min(1, 1/\lambda) \sum_{u \in V} \pi_u \left[ \pi_u + \sum_{v \neq u} \mathbb{E}|Z_{uv} - Z_v| \right].$$
Applications

Proposition (Binomial approximation)

One has for all $n, \lambda \leq n$:
$$d_{var}(\text{Bin}(n, \lambda/n), \text{Poisson}(\lambda)) \leq 2 \min(1, \lambda) \frac{\lambda}{n}$$
Applications

Proposition (Binomial approximation)
One has for all $n, \lambda \leq n$:
$$d_{\text{var}}(\text{Bin}(n, \lambda/n), \text{Poisson}(\lambda)) \leq 2 \min(1, \lambda) \frac{\lambda}{n}$$

Proposition (Isolated nodes)
In $G(n, p)$ with $np = \ln(n) + c$, noting $\lambda = n(1 - p)^{n-1} \sim e^{-c}$ and $X$: number of isolated nodes, then
$$d_{\text{var}}(X, \text{Poisson}(\lambda)) \leq 2\lambda[1/n + p/(1 - p)] = O(\ln(n)/n)$$
Hence, $\lim_{n \to \infty} P(X = 0) = e^{-e^{-c}}$
Stein-Chen method – proof arguments

Fact: for each $\lambda > 0, A \subset \mathbb{N}$, function $f : \mathbb{N} \to \mathbb{R}$ defined by

$$f(0) = 0, \ \lambda f(j + 1) - j \cdot f(j) = \mathbb{I}_A(j) - \text{Poi}_\lambda(A), \ j \in \mathbb{N}$$

is $\min(1, \lambda^{-1})$–Lipschitz
Stein-Chen method – proof arguments

Fact: for each $\lambda > 0, A \subset \mathbb{N}$, function $f : \mathbb{N} \to \mathbb{R}$ defined by

$$f(0) = 0, \lambda f(j + 1) - j \cdot f(j) = \mathbb{1}_A(j) - \text{Poi}_\lambda(A), j \in \mathbb{N}$$

is $\min(1, \lambda^{-1})$–Lipschitz

Write

$$|\mathbb{P}(X \in A) - \text{Poi}_\lambda(A)| = |\mathbb{E}[\lambda f(X + 1) - Xf(X)]|$$
**Stein-Chen method – proof arguments**

**Fact:** for each $\lambda > 0$, $A \subset \mathbb{N}$, function $f : \mathbb{N} \rightarrow \mathbb{R}$ defined by

$$f(0) = 0, \quad \lambda f(j + 1) - j \cdot f(j) = \mathbb{I}_A(j) - \text{Poi}_\lambda(A), \quad j \in \mathbb{N}$$

is $\min(1, \lambda^{-1})$–Lipschitz

Write

$$|\mathbb{P}(X \in A) - \text{Poi}_\lambda(A)| = |\mathbb{E}[\lambda f(X + 1) - X f(X)]|$$

$$= \left| \sum_{u \in V} \pi_u \mathbb{E} \left[ f(X + 1) - f(1 + \sum_{v \neq u} Z_{uv}) \right] \right|$$
Stein-Chen method – proof arguments

**Fact:** for each $\lambda > 0$, $A \subset \mathbb{N}$, function $f : \mathbb{N} \to \mathbb{R}$ defined by

$$f(0) = 0, \quad \lambda f(j + 1) - j \cdot f(j) = \mathbb{I}_A(j) - \text{Poi}_\lambda(A), \quad j \in \mathbb{N}$$

is $\min(1, \lambda^{-1})$–Lipschitz

Write

$$|\mathbb{P}(X \in A) - \text{Poi}_\lambda(A)| = |\mathbb{E}[\lambda f(X + 1) - Xf(X)]|$$

$$= \left| \sum_{u \in V} \pi_u \mathbb{E} \left[ f(X + 1) - f(1 + \sum_{v \neq u} Z_{uv}) \right] \right|$$

$$\leq \sum_{u \in V} \pi_u \min(1, \lambda^{-1}) \mathbb{E} \left| \sum_{v \in V} Z_v - \sum_{v \neq u} Z_{uv} \right|$$
Stein-Chen method – proof arguments

**Fact:** for each $\lambda > 0$, $A \subset \mathbb{N}$, function $f : \mathbb{N} \to \mathbb{R}$ defined by

$$f(0) = 0, \quad \lambda f(j + 1) - j \cdot f(j) = \mathbb{I}_A(j) - \text{Poi}_\lambda(A), \quad j \in \mathbb{N}$$

is $\min(1, \lambda^{-1})$–Lipschitz

Write

$$|\mathbb{P}(X \in A) - \text{Poi}_\lambda(A)| = |\mathbb{E}[\lambda f(X + 1) - Xf(X)]|$$

$$= \left| \sum_{u \in V} \pi_u \mathbb{E} \left[ f(X + 1) - f(1 + \sum_{v \neq u} Z_{uv}) \right] \right|$$

$$\leq \sum_{u \in V} \pi_u \min(1, \lambda^{-1}) \mathbb{E} \left| \sum_{v \in V} Z_v - \sum_{v \neq u} Z_{uv} \right|$$

$$\leq \sum_{u \in V} \pi_u \left[ \pi_u + \sum_{v \neq u} \mathbb{E}|Z_v - Z_{uv}| \right]$$
Connectivity – final arguments

Let $A_k = \{\exists$ connected component of size $k\}$. By union bound, for $p = \Theta(\ln(n)/n)$,

$\mathbb{P}(A_2) \leq \binom{n}{2} p(1 - p)^{2(n-2)} \leq O(p) = o(1)$
Connectivity – final arguments

Let $\mathcal{A}_k = \{\exists$ connected component of size $k\}$. By union bound, for $p = \Theta(\ln(n)/n)$,

$$\mathbb{P}(\mathcal{A}_2) \leq \binom{n}{2} p(1 - p)^2(n-2) \leq O(p) = o(1)$$

Similarly for $k \leq n/2$, $\mathbb{P}(\mathcal{A}_k) \leq \binom{n}{k} T_k p^{k-1}(1 - p)^k(n-k)$

where $T_k$: number of trees on $[k]$. 

Cayley’s theorem: $T_k = k^{k-2}$. 

Hence

$$\mathbb{P}(\mathcal{A}_k) \leq \binom{n}{k} k^{k-2} p^{k-1}(1 - p)^k(n-k) \leq \frac{n^k}{k!} k^{k-2} p^{k-1} e^{-pk/2} \leq \frac{1}{p} \frac{k}{k+1} \sqrt{k e^k (1+\ln(np)-np/2)}$$

Conclusion

$$\mathbb{P}(\bigcup_{2 \leq k \leq n/2} \mathcal{A}_k) \leq \sum_{2 \leq k \leq n/2} \mathbb{P}(\mathcal{A}_k) \to 0 \text{ as } n \to \infty$$
Let \( A_k = \{ \exists \text{ connected component of size } k \} \).

By union bound, for \( p = \Theta(\ln(n)/n) \),
\[
\mathbb{P}(A_2) \leq \binom{n}{2} p(1 - p)^2(n-2) \leq O(p) = o(1)
\]

Similarly for \( k \leq n/2, \mathbb{P}(A_k) \leq \binom{n}{k} T_k p^{k-1}(1 - p)^{k(n-k)} \)
where \( T_k \): number of trees on \([k]\)

**Cayley’s theorem:** \( T_k = k^{k-2} \).
Connectivity – final arguments

Let \( \mathcal{A}_k = \{ \exists \text{ connected component of size } k \} \).

By union bound, for \( p = \Theta(\ln(n)/n) \),

\[
P(\mathcal{A}_2) \leq \binom{n}{2} p(1 - p)^2(n-2) \leq O(p) = o(1)
\]

Similarly for \( k \leq n/2 \), \( P(\mathcal{A}_k) \leq \binom{n}{k} T_k p^{k-1}(1 - p)^{k(n-k)} \)

where \( T_k \): number of trees on \([k]\)

Cayley’s theorem: \( T_k = k^{k-2} \). Hence

\[
P(\mathcal{A}_k) \leq \binom{n}{k} k^{k-2} p^{k-1}(1 - p)^{k(n-k)}
\leq \frac{n^k}{k!} k^{k-2} p^{k-1} e^{-pkn/2}
\leq \frac{1}{p \sqrt{k}} e^{k(1+\ln(np) - np/2)}
\]
Connectivity – final arguments

Let $A_k = \{\exists$ connected component of size $k\}$. By union bound, for $p = \Theta(\ln(n)/n)$,
$$\mathbb{P}(A_2) \leq \binom{n}{2} p(1 - p)^{2(n-2)} \leq O(p) = o(1)$$

Similarly for $k \leq n/2$, $\mathbb{P}(A_k) \leq \binom{n}{k} T_k p^{k-1}(1 - p)^{k(n-k)}$ where $T_k$: number of trees on $[k]

Cayley’s theorem: $T_k = k^{k-2}$. Hence

$$\mathbb{P}(A_k) \leq \binom{n}{k} k^{k-2} p^{k-1}(1 - p)^{k(n-k)} \\
\leq \frac{n^k}{k!} k^{k-2} p^{k-1} e^{-pkn/2} \\
\leq \frac{1}{p} \frac{1}{k^2 \sqrt{k}} e^{k(1+\ln(np)-np/2)}$$

Conclusion $\mathbb{P}(\bigcup_{2 \leq k \leq n/2} A_k) \leq \sum_{2 \leq k \leq n/2} \mathbb{P}(A_k) \to 0$ as $n \to \infty$ follows.
Takeaway messages

- Connectivity of Erdős-Rényi graphs informs behaviour of SIR epidemics on complete graph

- Emergence of giant component of size $n(1 - p_{ext})$ as average degree crosses critical value 1

- Full connectivity for average degree $\ln(n) + O(1)$

- Proof techniques: branching process approximation, Chernoff bounds; First and second moment methods; Poisson approximation via Stein-Chen method