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

Laurent Massoulié

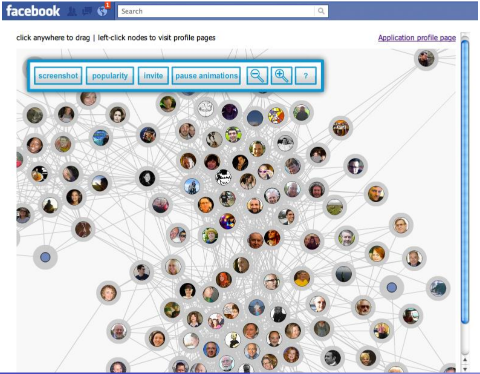
Inria

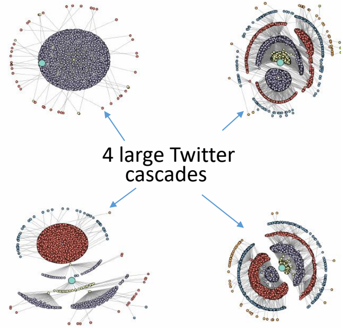
January 20, 2021

Viral propagation of information and "information cascades"

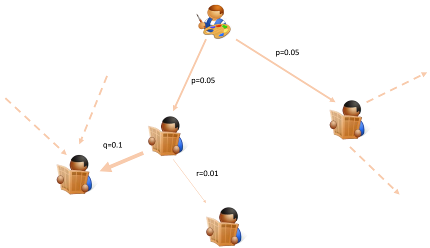
Propagation on underlying graph (e.g. facebook's "friendship graph", or Twitter's "follower-followee" directed graph)

→ Epidemic models to understand viral propagation (and guide viral marketing strategies)





The Independent Cascade, or Susceptible-Infective-Removed (SIR) epidemics model



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

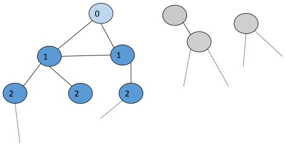
- Special case: complete graph on $i \in [n]$ and homogeneous infection probabilities $p_{ij} \equiv p$
- Associated model: Erdős-Rényi random graph $\mathcal{G}(n, p)$: undirected graph on node set $[n]$. Edge (i, j) present iff $\xi_{ij} = 1$ where $\{\xi_{ij}\}_{i<j}$: i.i.d., Bernoulli (p)

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 $\mathcal{G}(n, p)$: 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 ξ_{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 $\mathcal{G}(n,p)$: 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 ξ_{ij} to determine if when the first of i and j gets infected, it infects the other



$\Rightarrow$ For initial set X_0 of infective nodes at time 0, i infected at time t iff $d_G(X_0, i) = t$
Set of nodes eventually infected: $\cup_{i \in X_0} \Gamma(i)$ where $\Gamma(i)$: graph's connected component including i

Outline

Seminal results by Erdős and Rényi (1959-1960)

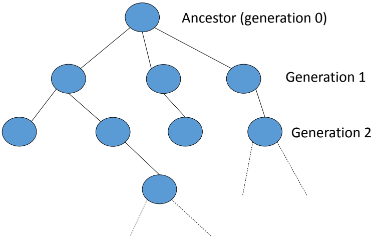
- First phase transition: emergence of giant component

  Tools: branching processes & Chernoff's inequality

- Second phase transition: emergence of connectivity

  Tools: 1st and 2nd moment methods; Poisson approximation

Towards Susceptible-Infective-Removed (SIR) epidemics: Galton-Watson branching process (1873)



Offspring distribution $\{p_k\}_{k\in\mathbb{N}}$
Z_k number of individuals per generation:
$Z_0 = 1, Z_k = \sum_{m=1}^{Z_{k-1}} X_{m,k}$ where $\{X_{m,k}\}_{m,k\geq 0}$: 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_{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* $p_{ext} = 1$
If $\mu = 1$ *and* $p_0 > 0$ *then* $p_{ext} = 1$
If $\mu > 1$ *then* $p_{ext} < 1$

Theorem

Extinction probability p_{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* $p_{ext} = 1$
If $\mu = 1$ *and* $p_0 > 0$ *then* $p_{ext} = 1$
If $\mu > 1$ *then* $p_{ext} < 1$

Proof: $\{Z_k = 0\} \nearrow \{\text{Extinction}\}$; $\mathbb{P}(Z_k = 0) = \phi_k(0)$ 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 ϕ and $\mathbb{P}(Z_0 = 0) = 0$, sequence increases to (necessarily smallest) fixed point.

Theorem

Extinction probability p_{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* $p_{ext}=1$
If $\mu=1$ *and* $p_0>0$ *then* $p_{ext}=1$
If $\mu>1$ *then* $p_{ext}<1$

Proof: $\{Z_k=0\}\nearrow\{\text{Extinction}\}$; $\mathbb{P}(Z_k=0)=\phi_k(0)$ 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 ϕ and $\mathbb{P}(Z_0=0)=0$, sequence increases to (necessarily smallest) fixed point.
μ: slope of ϕ at 1^-. By convexity of ϕ, only fixed point: 1 if $\mu<1$
By continuity of ϕ, $\exists$ fixed point <1 if $\mu>1$
For $\mu=1$, if $p_0>0$ then ϕ strictly convex hence only fixed point: 1; if $p_0=0$ then $p_{ext}=0$

Theorem

Extinction probability p_{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* $p_{ext} = 1$
If $\mu = 1$ *and* $p_0 > 0$ *then* $p_{ext} = 1$
If $\mu > 1$ *then* $p_{ext} < 1$

Proof: $\{Z_k = 0\} \nearrow \{\text{Extinction}\}$; $\mathbb{P}(Z_k = 0) = \phi_k(0)$ 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 ϕ and $\mathbb{P}(Z_0 = 0) = 0$, sequence increases to (necessarily smallest) fixed point.
μ: slope of ϕ at 1^-. By convexity of ϕ, only fixed point: 1 if $\mu < 1$
By continuity of ϕ, $\exists$ fixed point < 1 if $\mu > 1$
For $\mu = 1$, if $p_0 > 0$ then ϕ strictly convex hence only fixed point: 1; if $p_0 = 0$ then $p_{ext} = 0$
Fundamental example of **phase transition**
Special case $X \sim \text{Poisson}(\mu)$: $p_{ext} = e^{-\mu(1-p_{ext})}$

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)

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

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-Tchebitchev's inequality**: random variable $X \in \mathbb{R}$: $\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \mathrm{Var}(X)/a^2$
- **Exponential version**: for $\theta > 0$, $\mathbb{P}(X \geq t) \leq \mathbb{E}(e^{\theta X})e^{-\theta t}$ 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, $\mathbb{P}(\sum_{s=1}^{t} X_s \geq at) \leq e^{-th(a)}$ *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 $\exists \epsilon > 0 : \mathbb{E}e^{\epsilon X_1} < +\infty$

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.

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 $\exists \epsilon > 0 : \mathbb{E}e^{\epsilon X_1} < +\infty$

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 $\mathcal{G}(n,p)$

Theorem

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

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

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

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

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

where p_{ext}: extinction probability of Poisson (λ) 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 $1 - p_{ext}$, epidemics started from randomly selected node reaches $n[1 - p_{ext} + o(1)]$ 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 $\mathcal{A}_0 = \{i_0\}$ and killed set $\mathcal{B}_0 = \emptyset$
- At time t pick $j_t \in \mathcal{A}_{t-1}$, kill it and activate its neighbours not yet activated (set $\mathcal{D}_t$)
  $\Rightarrow \mathcal{A}_t = \mathcal{A}_{t-1} \setminus \{j_t\} \cup \mathcal{D}_t,\ \mathcal{B}_t = \mathcal{B}_{t-1} \cup \{j_t\}$

Sub-critical regime

- Exploration of connected component $\Gamma(i_0)$: initialized with active set $\mathcal{A}_0 = \{i_0\}$ and killed set $\mathcal{B}_0 = \emptyset$
- At time t pick $j_t \in \mathcal{A}_{t-1}$, kill it and activate its neighbours not yet activated (set $\mathcal{D}_t$)
  $\Rightarrow \mathcal{A}_t = \mathcal{A}_{t-1} \setminus \{j_t\} \cup \mathcal{D}_t, \ \mathcal{B}_t = \mathcal{B}_{t-1} \cup \{j_t\}$
- Notation: $A_t = |\mathcal{A}_t|, \ D_t = |\mathcal{D}_t| \ \Rightarrow A_t = 1 - t + D_1 + \cdots + D_t$

Sub-critical regime

- Exploration of connected component $\Gamma(i_0)$: initialized with active set $\mathcal{A}_0 = \{i_0\}$ and killed set $\mathcal{B}_0 = \emptyset$
- At time t pick $j_t \in \mathcal{A}_{t-1}$, kill it and activate its neighbours not yet activated (set $\mathcal{D}_t$)
  $\Rightarrow \mathcal{A}_t = \mathcal{A}_{t-1} \setminus \{j_t\} \cup \mathcal{D}_t,\ \mathcal{B}_t = \mathcal{B}_{t-1} \cup \{j_t\}$
- Notation: $A_t = |\mathcal{A}_t|,\ D_t = |\mathcal{D}_t|\ \Rightarrow A_t = 1 - t + D_1 + \cdots + D_t$
- Conditionally on $\mathcal{F}_{t-1} = \sigma(A_1, \ldots, A_{t-1})$,
  $D_t \sim \mathrm{Bin}(p, n - 1 - D_1 - \cdots - D_{t-1})$

Sub-critical regime

- Exploration of connected component $\Gamma(i_0)$: initialized with active set $\mathcal{A}_0 = \{i_0\}$ and killed set $\mathcal{B}_0 = \emptyset$
- At time t pick $j_t \in \mathcal{A}_{t-1}$, kill it and activate its neighbours not yet activated (set $\mathcal{D}_t$)
  $\Rightarrow \mathcal{A}_t = \mathcal{A}_{t-1} \setminus \{j_t\} \cup \mathcal{D}_t,\ \mathcal{B}_t = \mathcal{B}_{t-1} \cup \{j_t\}$
- Notation: $A_t = |\mathcal{A}_t|,\ D_t = |\mathcal{D}_t|\ \Rightarrow A_t = 1 - t + D_1 + \cdots + D_t$
- 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 $h(x) = \lambda h_1(x/\lambda),\ h_1(x) = x\ln(x) - x + 1$: Chernoff's exponent for Poisson (λ) 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 (λ) 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 $\lim_{n\to\infty} \mathbb{P}(C \le k) = \mathbb{P}(Z \le k) \le p_{ext}$
where Z*: total population of Poisson* (λ) *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 λ, graph disconnected $\Rightarrow$ Under what regime is graph connected?

Connectivity

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

Theorem

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

Connectivity

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

Theorem

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

Corollary

If $np - \ln(n) \to +\infty$, then $\lim_{n\to\infty} \mathbb{P}(\mathcal{G}(n,p)$ connected$) = 1$
If $np - \ln(n) \to -\infty$, then $\lim_{n\to\infty} \mathbb{P}(\mathcal{G}(n,p)$ 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 $\lim_{n\to\infty} \mathbb{P}(\mathcal{A}) = e^{-e^{-c}}$ where $\mathcal{A} = \{\text{no isolated vertices in } \mathcal{G}(n,p)\}$

Proof strategy

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

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

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

Proof strategy

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

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

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

- Use bounds

$$\mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{B}) \le \mathbb{P}(\mathcal{G}(n,p) \text{ connected}) = \mathbb{P}(\mathcal{A} \cap \overline{\mathcal{B}}) \le \mathbb{P}(\mathcal{A})$$

Basic tools: the first and second moment methods

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

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

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 $\mathcal{G}(n, p)$ 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 $\mathcal{G}(n, p)$ if $\lim_{n \to \infty}[np - \ln(n)] = +\infty$.

Second moment method: $\mathbb{P}(\forall u \in V, Z_u = 0) = \mathbb{P}(X = 0) \leq \frac{\mathrm{Var}(X)}{\mathbb{E}(X)^2}$. Hence if $\mathrm{Var}(X) = o(\mathbb{E}(X)^2)$, 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 $\mathcal{G}(n, p)$ if $\lim_{n \to \infty}[np - \ln(n)] = +\infty$.

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

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

Variation distance

Definition

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

Variation distance

Definition

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

Alternative characterization: if μ, ν admit densities $\frac{d\mu}{d\pi}, \frac{d\nu}{d\pi}$ with respect to measure π (e.g., $\pi = \mu + \nu$) then
$d_{var}(\mu, \nu) = \int_\Omega \left| \frac{d\mu}{d\pi} - \frac{d\nu}{d\pi} \right| d\pi$
In particular for $\Omega = \mathbb{N}$ and $\pi = \sum_{n \in \mathbb{N}} \delta_n$, $d_{var}(\mu, \nu) = \sum_{n \in \mathbb{N}} |\mu_n - \nu_n|$

Variation distance

Definition

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

Alternative characterization: if μ, ν admit densities $\frac{d\mu}{d\pi}, \frac{d\nu}{d\pi}$ with respect to measure π (e.g., $\pi = \mu + \nu$) then
$d_{var}(\mu, \nu) = \int_\Omega \left| \frac{d\mu}{d\pi} - \frac{d\nu}{d\pi} \right| d\pi$
In particular for $\Omega = \mathbb{N}$ and $\pi = \sum_{n \in \mathbb{N}} \delta_n$, $d_{var}(\mu, \nu) = \sum_{n \in \mathbb{N}} |\mu_n - \nu_n|$

Definition

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

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

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:

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

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

Proposition (Isolated nodes)

In $\mathcal{G}(n,p)$ with $np = \ln(n) + c$, noting $\lambda = n(1-p)^{n-1} \sim e^{-c}$ and X: number of isolated nodes, then

$$d_{var}(X, \text{Poisson}(\lambda)) \leq 2\lambda[1/n + p/(1-p)] = O(\ln(n)/n)$$

Hence, $\lim_{n\to\infty} \mathbb{P}(X=0) = e^{-e^{-c}}$

Stein-Chen method – proof arguments

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

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

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

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

is $\min(1, \lambda^{-1})$–Lipschitz

Write

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

Stein-Chen method – proof arguments

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

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

is $\min(1, \lambda^{-1})$–Lipschitz

Write

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

Stein-Chen method – proof arguments

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

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

is $\min(1, \lambda^{-1})$–Lipschitz

Write

$$\begin{aligned} |\mathbb{P}(X \in A) - \mathrm{Poi}_\lambda(A)| &= |\mathbb{E}[\lambda f(X+1) - Xf(X)]| \\ &= \left| \sum_{u \in V} \pi_u \mathbb{E}\left[f(X+1) - f(1 + \sum_{v \neq u} Z_{uv}) \right] \right| \\ &\leq \sum_{u \in V} \pi_u \min(1, \lambda^{-1}) \mathbb{E} \left| \sum_{v \in V} Z_v - \sum_{v \neq u} Z_{uv} \right| \end{aligned}$$

Stein-Chen method – proof arguments

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

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

is $\min(1, \lambda^{-1})$–Lipschitz

Write

$$\begin{aligned}
|\mathbb{P}(X \in A) - \mathrm{Poi}_\lambda(A)| &= |\mathbb{E}[\lambda f(X+1) - Xf(X)]| \\
&= \left| \sum_{u \in V} \pi_u \mathbb{E}\left[f(X+1) - f(1 + \sum_{v \neq u} Z_{uv}) \right] \right| \\
&\leq \sum_{u \in V} \pi_u \min(1, \lambda^{-1}) \mathbb{E} \left| \sum_{v \in V} Z_v - \sum_{v \neq u} Z_{uv} \right| \\
&\leq \sum_{u \in V} \pi_u \left[\pi_u + \sum_{v \neq u} \mathbb{E}|Z_v - Z_{uv}| \right]
\end{aligned}$$

Connectivity – final arguments

Let $\mathcal{A}_k = \{\exists$ connected component of size $k\}$.
By union bound, for $p = \Theta(\ln(n)/n)$,
$\mathbb{P}(\mathcal{A}_2) \leq \binom{n}{2} p(1-p)^{2(n-2)} \leq O(p) = o(1)$

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 $[k]$

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: $T_k = k^{k-2}$.

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

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

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

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

Conclusion $\mathbb{P}(\cup_{2\leq k\leq n/2}\mathcal{A}_k) \leq \sum_{2\leq k\leq n/2} \mathbb{P}(\mathcal{A}_k) \to 0$ 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