Some approximation results for subcritical Erdős-Rényi random graphs via Stein's method

Fraser Daly

Heriot–Watt University

Applied Probability Workshop, 27 August 2020

Joint work with Seva Shneer (Heriot–Watt)
https://arxiv.org/abs/1912.03219

Outline

- Subcritical Erdős-Rényi random graphs
- Geometric approximation for typical path lengths in subcritical Erdős-Rényi graphs
- Borel approximation for typical component size in subcritical Erdős-Rényi graphs

Erdős-Rényi random graphs

Consider the Erdős-Rényi graph $G(n, \lambda/n)$ with n vertices. Each pair of vertices are connected by an edge independently with probability λ/n, for some parameter $\lambda > 0$.

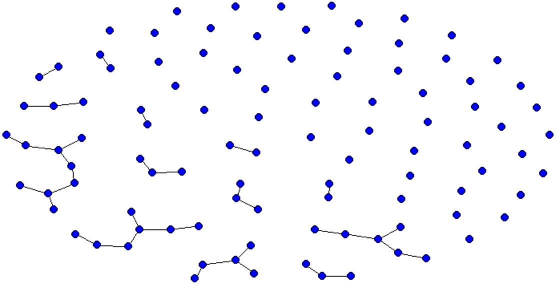

Figure: $n = 100$, $\lambda = 0.8$

Erdős-Rényi random graphs

Consider the Erdős-Rényi graph $G(n, \lambda/n)$ with n vertices. Each pair of vertices are connected by an edge independently with probability λ/n, for some parameter $\lambda > 0$.

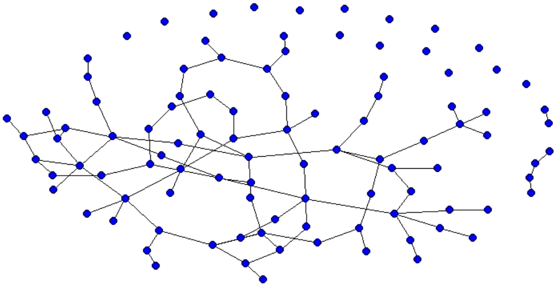

Figure: $n = 100$, $\lambda = 2$

Erdős-Rényi random graphs

Consider the sizes of components (sets of connected vertices) in $G(n, \lambda/n)$. Asymptotically (as $n \to \infty$):

- When $\lambda < 1$ (the *subcritical* case), with high probability all components of the graph are small, of order at most $\log(n)$; two uniformly chosen vertices are likely to be in different components.
- When $\lambda = 1$ (the *critical* case), with high probability there are many components with a size of order $n^{2/3}$.
- When $\lambda > 1$ (the *supercritical* case), with high probability there is a single giant component with a non-zero proportion of the vertices, and all other components are of order at most $\log(n)$.

Erdős-Rényi random graphs

Consider the sizes of components (sets of connected vertices) in $G(n, \lambda/n)$. Asymptotically (as $n \to \infty$):

- When $\lambda < 1$ (the *subcritical* case), with high probability all components of the graph are small, of order at most $\log(n)$; two uniformly chosen vertices are likely to be in different components.

Our interest in the subcritical case.

Subcritical Erdős-Rényi random graphs

For the subcritical graph $G(n, \lambda/n)$ for $0 < \lambda < 1$, we define

- L: the length of the shortest path between vertices 1 and 2 (if they are connected; ∞ otherwise).
- C: the number of vertices in the same component as vertex 1 (i.e., the typical component size).

We will derive explicit error bounds in the approximation of $L|L < \infty$ by a geometric random variable and in the approximation of C by a Borel random variable.

Geometric approximation for $L|L < \infty$

With high probability, L is infinite. Conditioning on vertices 1 and 2 being in the same component of the graph, $L|L < \infty$ is known to be asymptotically $\mathrm{Geom}(1-\lambda)$, as $n \to \infty$. See Katzav, Biham and Hartmann (2018).

We write $X \sim \mathrm{Geom}(p)$ if $\mathbb{P}(X = j) = p(1-p)^{j-1}$ for $j = 1, 2, \ldots$.

We can explicitly calculate

$$\mathbb{P}(L = 1|L < \infty) = \frac{\mathbb{P}(L = 1)}{\mathbb{P}(L < \infty)} = (1-\lambda) + \frac{\lambda}{n}.$$

Geometric approximation for $L|L < \infty$

We give an explicit error bound in total variation distance:

$$d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) = \sup_{A \subseteq \mathbb{N}} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|$$
$$= \frac{1}{2} \sup_{\substack{h:\mathbb{N}\mapsto\mathbb{R} \\ \|h\|\leq 1}} |\mathbb{E}h(X) - \mathbb{E}h(Y)| = \inf_{(X,Y)} \mathbb{P}(X \neq Y)\,,$$

where the infimum is taken over all couplings of (X, Y).

To do this we use Stein's method for geometric approximation, as developed by Peköz (1996). This is based on the observation that $X \sim \text{Geom}(p)$ if and only if

$$X + 1 \stackrel{d}{=} X|X > 1\,,$$

where "$\stackrel{d}{=}$" denotes equality in distribution.

Geometric approximation for $L|L<\infty$

Letting $p_n = \mathbb{P}(L=1|L<\infty)$ and $Y_n \sim \mathrm{Geom}(p_n)$, define f_A to be the solution of

$$I(k \in A) - \mathbb{P}(Y_n \in A) = (1-p_n)f_A(k+1) - f_A(k)\,,$$

for $A \subseteq \mathbb{N}$. We have that $\sup_{j,k}|f_A(j)-f_A(k)| \leq \frac{1}{p_n}$ for each A.

This 'Stein equation' is motivated by the fact that when replacing k by $L|L<\infty$, taking absolute values and taking the supremum over $A \subseteq \mathbb{N}$:

- The LHS is the total variation distance between $L|L<\infty$ and Y_n.
- The RHS compares $L+1|L<\infty$ with $L|1<L<\infty$.

Geometric approximation for $L|L < \infty$

Then

$$\begin{aligned}
&d_{TV}(\mathcal{L}(L|L<\infty), \mathrm{Geom}(p_n)) \\
&\quad = \sup_{A\subseteq\mathbb{N}} |(1-p_n)\mathbb{E}[f_A(L+1)|L<\infty] - \mathbb{E}[f_A(L)|L<\infty]| \\
&\quad = (1-p_n) \sup_{A\subseteq\mathbb{N}} |\mathbb{E}[f_A(L+1)|L<\infty] - \mathbb{E}[f_A(L)|1<L<\infty]| \\
&\quad \leq \frac{1-p_n}{p_n} d_{TV}(\mathcal{L}(L+1|L<\infty), \mathcal{L}(L|1<L<\infty)) \\
&\quad \leq \frac{\lambda}{1-\lambda} d_{TV}(\mathcal{L}(L+1|L<\infty), \mathcal{L}(L|1<L<\infty)) \,.
\end{aligned}$$

Geometric approximation for $L|L < \infty$

A realization of $L|1 < L < \infty$ gives us a shortest path (of length at least two) from vertex 1 to vertex 2. Give the penultimate vertex on this path the label 3. The path from 1 to 3 gives us a realization of $L|L < \infty$, up to the fact that vertex 3 is not (quite) uniformly chosen. Hence,

$$d_{TV}(\mathcal{L}(L|1 < L < \infty), \mathcal{L}(L+1|L < \infty)) \leq \frac{1}{n-1}\,.$$

If we want an explicit error bound for the approximation of $L|L < \infty$ by Geom$(1-\lambda)$, we can use the triangle inequality and a simple bound between two geometric distributions to get the following.

Theorem

$$d_{TV}(\mathcal{L}(L|L < \infty), Geom(1-\lambda)) \leq \frac{\lambda(2-\lambda+\lambda^2)}{(1-\lambda)^3(n-1)}\,.$$

Borel approximation for C

Asymptotically (as $n \to \infty$), C is known to have a Borel(λ) distribution. $Z \sim$ Borel(λ) satisfies

$$Z \stackrel{d}{=} 1 + \sum_{i=1}^{\xi} Z_i \,,$$

where $Z, Z_1, Z_2, \ldots$ are i.i.d. and $\xi \sim$ Po(λ) has a Poisson distribution.

Thus, Z represents the total progeny in a Galton–Watson process with Poisson offspring distribution. Its appearance as the limit of C is a consequence of the branching approximation for $G(n, \lambda/n)$.

Borel approximation for C

Asymptotically (as $n \to \infty$), C is known to have a Borel(λ) distribution. $Z \sim$ Borel(λ) satisfies

$$Z \stackrel{d}{=} 1 + \sum_{i=1}^{\xi} Z_i \,,$$

where $Z, Z_1, Z_2, \ldots$ are i.i.d. and $\xi \sim$ Po(λ) has a Poisson distribution.

Thus, Z represents the total progeny in a Galton-Watson process with Poisson offspring distribution. Its appearance as the limit of C is a consequence of the branching approximation for $G(n, \lambda/n)$.

We have that

$$\mathbb{P}(Z = j) = \frac{e^{-\lambda j}(\lambda j)^{j-1}}{j!} \,,$$

for $j = 1, 2, \ldots$ and that $\mathbb{E}Z = \frac{1}{1-\lambda}$.

Borel approximation for C

For any non-negative integer valued random variable X (with $\mathbb{E}X > 0$), we can define $X^\star$, the size-biased version of X, with

$$\mathbb{P}(X^\star = j) = \frac{j\mathbb{P}(X = j)}{\mathbb{E}X},$$

for $j = 1, 2, \ldots$.

Using rules for size biasing random sums [see Arratia, Goldstein and Kochman (2019)], we can use the random sum representation of $Z \sim \text{Borel}(\lambda)$ to get that

$$Z^\star \stackrel{d}{=} (1 - I)Z + I(Z + Z^\star),$$

where I is independent of all else with

$$\mathbb{P}(I = 1) = 1 - \mathbb{P}(I = 0) = \lambda.$$

Borel approximation for C

By comparing C with the total number of infected individuals in a Reed–Frost epidemic model, results of Ball and Donnelly (1995) give an upper bound of order $O(n^{-1})$ on $d_{TV}(\mathcal{L}(C), \text{Borel}(\lambda))$.

Here we will analyse this problem using Stein's method, based on the above characterisation of the Borel distribution. Unfortunately we will only obtain a bound of order $O(\frac{\log(n)}{n})$, and for a restricted range of values of λ. This should be thought of as a first attempt at using Stein's method for Borel approximation, which leaves open research questions we will highlight at the end.

Borel approximation for C

We construct a Stein equation that compares the distribution of $C^\star$ with $(1-I)C + I(Z + C^\star)$.

Let f_A be the solution of

$$I(k \in A) - \mathbb{P}(Z \in A) = (1-\lambda)(k-1)f_A(k) - \lambda(1-\lambda)k\sum_{i=1}^{\infty} f_A(i+k)\mathbb{P}(Z=i)\,,$$

where $Z \sim \text{Borel}(\lambda)$, so that

$$d_{TV}(\mathcal{L}(C), \text{Borel}(\lambda)) = \sup_{A\subseteq\mathbb{N}} \left|\mathbb{E}f_A(C^\star) - \mathbb{E}f_A((1-I)C + I(Z+C^\star))\right|.$$

Borel approximation for C

We can show that $\sup_k |f_A(k)| \leq (1-\lambda)^{-2}$ for each A, and hence

$$d_{TV}(\mathcal{L}(C), \mathcal{L}(Z)) \leq \frac{1}{(1-\lambda)^2} d_{TV}(\mathcal{L}(C^\star), \mathcal{L}((1-I)C + I(Z+C^\star))),$$

Writing $C = 1 + \sum_{j=2}^n I(\text{vertex } j \text{ is connected to vertex } 1)$, we can calculate

$$C^\star \stackrel{d}{=} (1-I')C + I'(C|L<\infty),$$

where $\mathbb{P}(I'=1) = \lambda - \frac{\lambda}{n}$.

Borel approximation for C

Coupling I and I' monotonically, and conditioning on their values, we thus get

$$d_{TV}(\mathcal{L}(C), \mathcal{L}(Z)) \leq \frac{\lambda}{(1-\lambda)^2}\left(d_{TV}(\mathcal{L}(C|L<\infty), \mathcal{L}(Z+C^\star)) + \frac{1}{n}\right).$$

Replacing the remaining Z on the RHS by C (using the triangle inequality), we get

$$\begin{aligned} d_{TV}&(\mathcal{L}(C), \mathcal{L}(Z)) \\ &\leq \frac{\lambda}{1-3\lambda+\lambda^2}\left(d_{TV}(\mathcal{L}(C|L<\infty), \mathcal{L}(C+C^\star)) + \frac{1}{n}\right), \end{aligned}$$

as long as $\lambda < \frac{1}{2}(3-\sqrt{5}) \approx 0.38$.

Borel approximation for C

There is another copy of I' 'hidden' in the $C^\star$ here: we 'match' this with an indicator $I(L > 1 | L < \infty)$ and get

$$d_{TV}(\mathcal{L}(C), \mathcal{L}(Z)) \\ \leq \frac{\lambda}{1 - 3\lambda + \lambda^2}\left(d_{TV}(\mathcal{L}(C|L=1), \mathcal{L}(C + \widetilde{C})) + \lambda\alpha_0 + \frac{1}{n}\right),$$

where $\widetilde{C}$ is an independent copy of C, and

$$\alpha_j = d_{TV}(\mathcal{L}(C|j+1 < L < \infty), \mathcal{L}(\widetilde{C} + C|j < L < \infty)).$$

By conditioning on the presence of an edge between vertices 1 and 2, we can bound

$$d_{TV}(\mathcal{L}(C|L=1), \mathcal{L}(C + \widetilde{C})) \leq \mathbb{P}(L < \infty) \leq \frac{\lambda}{(1-\lambda)n}.$$

Borel approximation for C

It remains only to bound α_0. By conditioning on the value of L,

$$\alpha_j \leq \theta_j \alpha_{j+1} + d_{TV}(\mathcal{L}(C|L=j+2), \mathcal{L}(\widetilde{C} + C|L=j+1)) + |\theta_j - \theta_{j+1}|\,,$$

where $\theta_j = \mathbb{P}(L > j+1 | j < L < \infty)$ and as above we can bound

$$d_{TV}(\mathcal{L}(C|L=j+2), \mathcal{L}(\widetilde{C} + C|L=j+1)) \leq \frac{\lambda(j+2)}{(1-\lambda)n}\,.$$

Applying this $m = O(\log(n))$ times to bound α_0 we get

$$\alpha_0 \leq \frac{\lambda}{(1-\lambda)n}\left(\sum_{j=0}^{m}(j+2)\Theta_j\right) = \sum_{j=0}^{m}|\theta_j - \theta_{j+1}|\Theta_j + \Theta_{m+1}\,,$$

where $\Theta_j = \mathbb{P}(L > j | L < \infty)$.

Borel approximation for C

We use our geometric approximation results from above to bound $\Theta_j = \mathbb{P}(L > j | L < \infty)$:

$$\lambda^j - \frac{a(\lambda)}{n-1} \leq \Theta_j \leq \lambda^j + \frac{a(\lambda)}{n-1} \,,$$

where $a(\lambda) = \frac{\lambda(2-\lambda+\lambda^2)}{(1-\lambda)^3}$. Similarly for $|\theta_j - \theta_{j+1}|$.

These give us an upper bound on α_0 of order $O\left(\frac{\log(n)}{n}\right)$, which may be combined with the above to obtain

$$d_{TV}(\mathcal{L}(C), \text{Borel}(\lambda)) \leq O\left(\frac{\log(n)}{n}\right) \,.$$

This is (slightly) worse than the bound $O(n^{-1})$ available from the results of Ball and Donnelly (1995).

Borel approximation for C

- We have an explicit choice for m:

$$m = \left\lfloor \frac{\log(n-1) - \log\log(n)}{-\log(\lambda)} - 1 \right\rfloor ,$$

  and a corresponding bound with an explicit (not too large) constant.

- We have assumed that $n \geq 19$ and $0 \leq m \leq n-4$, and that $\lambda < 0.38$. This final condition seems to be an artefact of the proof only (in particular, from the extra Z on the RHS of our Stein equation).

- Can we remove the condition $\lambda < 0.38$ and/or match the $O(n^{-1})$ bound using Stein's method? Is there a useful alternative Stein equation?

References

- R. Arratia, L. Goldstein and F. Kochman (2019). Size bias for one and all. *Probab. Surveys* **16**: 1–61.
- F. Ball and P. Donnelly (1995). Strong approximations for epidemic models. *Stoch. Proc. Appl.* **55**: 1–21.
- F. Daly and S. Shneer (2020). Stein's method for Borel approximation. Preprint. `arXiv:1912.03219`.
- E. Katzav, O. Biham and A. K. Hartmann (2018). Distribution of shortest path lengths in subcritical Erdős-Rényi networks. *Phys. Rev. E* 98: 012301.
- E. Peköz (1996). Stein's method for geometric approximation. *J. Appl. Prob.* **33**: 707–713.