Exploring Erdős-Rényi random graphs with IONTW*

Winfried Just† Hannah Callender‡ M. Drew LaMar§

January 8, 2016

In this module we explore in detail the distribution of the sizes of connected components of Erdős-Rényi random graphs and discover the reasons for the similarities and differences between disease transmission on Erdős-Rényi networks and complete graphs that were observed in the explorations of Module 6 of [2].

1 Introduction

As we explained in the brief overview of network-based models of transmission of infectious diseases at this web site[1], for most populations of hosts the actual contact network is not known, and we want to model it as a *random graph.* There are various constructions of such random graphs. They give us networks that usually share some, but not all properties of real contact networks. The most basic of these constructions gives *Erdős-Rényi random graphs,* named after the two Hungarian mathematicians who first systematically explored these graphs in the seminal paper [1]. These graphs serve as a benchmark against which all other constructions of random networks can be compared.

1.1 Definition of Erdős-Rényi random graphs

To construct such an Erdős-Rényi graph, we first decide on the set of nodes $\{1, 2, \ldots, N\}$. Then we list all edges $e_1, \ldots, e_{\frac{N(N-1)}{2}}$ of the *complete* graph K_N with N nodes and repeatedly toss a biased coin that comes up heads with probability p. We include edge e_ℓ as an actual edge of the *random* graph if, and only if, the coin comes up heads in toss number ℓ.

The mean degree $\langle k \rangle$ of the resulting graph will be approximately

$$\langle k \rangle \approx \lambda = p(N-1). \tag{1}$$

†Department of Mathematics, Ohio University, Athens, OH 45701 E-mail: mathjust@gmail.com

‡University of Portland E-mail: callende@up.edu

§The College of William and Mary E-mail: drew.lamar@gmail.com

[1] `http://www.ohio.edu/people/just/IONTW`

Thus by choosing a suitable value of the *connection probability* p, one can assure that the mean degree $\langle k \rangle$ of an Erdős-Rényi random graph will be close to the values that one might have estimated from data on real networks.

It will be more convenient if we think of Erdős-Rényi random graphs in terms of the parameter λ instead of the parameter p. The connection probability p can then be easily calculated as $p = \frac{\lambda}{N-1}$. The symbol $G_{ER}(N, \lambda)$ will denote *an* Erdős-Rényi random graph that is constructed with parameters N and λ.

Note that we used the indefinite article *an* in the previous sentence. $G_{ER}(N, \lambda)$ is not a uniquely determined; in fact, it could be *any* graph G with vertex set $V(G) = \{1, \ldots, N\}$. The symbol $G_{ER}(N, \lambda)$ only signifies that the graph is randomly drawn from a specific probability distribution. We will call a particular graph that has been constructed by the method described above *an instance of* $G_{ER}(N, \lambda)$. Note that (1) contains the symbol $\approx$. For a given *instance* of an Erdős-Rényi random graph we may not have exact equality $\langle k \rangle = \lambda$, but λ can be treated as the mean of $\langle k \rangle$ for all instances.

1.2 Properties of classes of networks

When we study random networks, we will no longer be able to definitely assert that a given random graph of a certain type has some property of interest to us. Instead, we will study the properties of a *class* of graphs.

The notion of a class of graphs takes some getting-used to. For a given value of the parameter λ, the construction of $G_{ER}(N, \lambda)$ defines a class of graphs that contains instances for arbitrarily large network sizes N. While for a given instance the mean degree $\langle k \rangle$ may differ from λ, when N is large, it will be very close to λ with probability that is very close to 1.

More precisely, for any fixed error bound $\varepsilon > 0$, the probability that the mean degree $\langle k \rangle$ of a randomly drawn instance $G_{ER}(N, \lambda)$ will differ from λ by more than ε will approach 0 as $N \to \infty$. Mathematicians say that the mean degree $\langle k \rangle$ of the actual instances approaches λ *asymptotically almost surely,* abbreviated *a.a.s.*

Thus all the properties of random graphs that we will be considering in this and subsequent modules are *average* properties. The best we can hope for is that some properties will hold a.a.s.

For another example, consider the degree distribution of $G_{ER}(N, \lambda)$. *On average,* this should be a binomial distribution with parameters N and $p = \frac{\lambda}{N-1}$, but for any given instance it may be slightly different. Binomial degree distributions are a bit cumbersome to work with. For large N and fixed λ it can be approximated reasonably well with a *Poisson distribution* with parameter λ, so that

$$q_k \approx \frac{\lambda^k e^{-\lambda}}{k!}, \tag{2}$$

where q_k is the fraction of nodes with degree k. Again, for any fixed error bound $\varepsilon > 0$ and nonnegative integer k, the probability that the proportion q_k of nodes with degree k in

randomly drawn instance $G_{ER}(N, \lambda)$ will differ from $\frac{\lambda^k e^{-\lambda}}{k!}$ by more than ε will approach 0 as $N \to \infty$. In other words, q_k will approach $\frac{\lambda^k e^{-\lambda}}{k!}$ a.a.s.

These probabilities q_k decrease rapidly to 0 as $k \to 0$. In particular, as $N \to \infty$, the expected maximum degree in $G_{ER}(N, \lambda)$ will a.a.s. be less than $\ln(N)$.

The phrase "asymptotically almost surely" also can be used for properties that are categorical rather than expressed in numerical values. For example, for certain types of random graphs it can be shown that they will a.a.s. be connected, which means that as $N \to \infty$ the probability of drawing a disconnected instance from this class with N nodes approaches zero. For other classes of random graphs it can be shown that a.a.s. they contain many different connected components. This does not imply even for very large N that we could not accidentally draw a graph from the former class that is disconnected or from the latter that is connected. It only implies that for very large N these events become very, very unlikely.

2 Exploring the connected components of Erdős-Rényi random graphs

Open IONTW, press **Defaults**, set the speed control slider to the extreme right, and use the following parameter settings.

model-time: Discrete
infection-prob: 1
end-infection-prob: 1
network-type → **Erdos-Renyi**
num-nodes: 300
lambda: 1.5
auto-set: Off

These parameter settings specify a next-generation SIR-model on an Erdős-Rényi network $G_{ER}(300, 1.5)$. Press **New** to look at the network. It is not possible to visually make out the connected components. But we can use the properties of the disease transmission model to visualize them: Since the probability b of an effective contact until the next time step, controlled by the input field **infection-prob**, is equal to 1, all nodes in the connected component of the index case j^* will eventually experience infection. The input setting **end-infection-prob** = 1 specifies a next-generation SIR-model in which all infectious nodes will get removed after exactly one time step and turn grey. If initially there is exactly one index case j^* in an otherwise susceptible population, all nodes outside of the connected component of j^* will remain green, and the connected component of j^* will show up in grey at the end of the simulation.

To see how this works, press **Set** to introduce one infectious node, and then **Go**. Repeat about 10 times for this network using **Reset** and then **Set**. This will keep the network fixed, but will change the initially infectious node. Record the approximate sizes of the connected

component by moving your mouse over the relevant part of the grey curve in the **Disease Prevalence** plot.

Exercise 1 *What do you observe? Do you get connected components with a range of different sizes? If the component is large, is it always the same one? How can you tell from the plot?*

The results may look puzzling. Repeat 10 more times, but look at a different instance $G_{ER}(300, 1.5)$ each time by pressing **New** instead of **Reset** before pressing **Set**.

Exercise 2 *In what respect are the results similar to the ones of the previous exercise; in what respect are they different?*

This is interesting. It appears that there is always one very large component in addition to many small ones.

Let us try to confirm the results we have discovered so far by running a large batch of simulations instead of looking at a few instances.

Switch **auto-set: On**

With the current parameter settings, define and run a batch processing experiment by using the template given at this web site[2] and the following specifications:

Define a **New** experiment.
Repetitions: 100
Measure runs using these reporters:
`count turtles with [removed?]`
Setup commands:
`new-network`

Exercise 3 *After the experiment is completed, open and analyze your output files. The column with the header* `count turtles with [removed?]` *reports the sizes of the connected component of the initially infectious node. Try to make out a distinctive gap between small and large components that were reported. Then record the maximum size of the observed small components and the mean size of the observed large components. Express these numbers as fractions of the total population size. Do your results confirm the preliminary observations that you made in the previous exercises?*

Now let us put our observation into the context of mathematical theory. Let us assume that the mean degree λ remains fixed, but N is allowed to be arbitrarily large. It can be shown that for $\lambda > 1$ there exist positive constants $\varrho(\lambda)$ and $c_{small}(\lambda)$ and such that a.a.s. (asymptotically almost surely) the proportion of nodes in the largest connected component of $G_{ER}(N, \lambda)$ will approach ϱ, while the size of all other components will not

[2] `http://www.ohio.edu/people/just/IONTW`

exceed $c_{small}(\lambda)\ln(N)$. In contrast, when $\lambda < 1$, a.a.s. all connected components of the graph $E(G_{ER}(N,\lambda)$ will have size $\leq c_{small}(\lambda)\ln(N)$.

For large N, the value of $c_{small}(\lambda)\ln(N)$ is much, much smaller than N. In mathematical language, $\lim_{N\to\infty}\frac{c_{small}(\lambda)\ln(N)}{N} = 0$. This implies that for $\lambda < 1$ we would expect that each connected component of $G_{ER}(N,\lambda)$ comprises only a very small proportion of the set of all nodes. The same will be true for all components other than the largest one when $\lambda > 1$. In particular, for all choices of λ the graphs $E(G_{ER}(N,\lambda)$ will a.a.s. contain many small connected components; in particular, they will a.a.s. be disconnected.

In the literature, the large connected component that we observed for $\lambda > 1$ is usually referred to as the *giant component.* Technically, a class of random graphs is said to *contain giant components a.a.s.* if for some fixed constant $\Theta > 0$ the probability that a graph from this class contains at least one connected component of size $\geq \Theta N$ approaches 1 as the size N of its vertex set increases without bound.

This definition is a mouthful. Let us take a close look at it. The definition assumes a fixed proportion Θ. It does not matter how close Θ is to 0; we only require that it is fixed and positive. For large N, the class is assumed to contain many networks of size N. Now fix N. For some of the graphs of size N in the class, the proportion of nodes in the largest connected component may be less than Θ. But if we choose N sufficiently large, *most* of these graphs will have a component of size $\geq \Theta N$. For large N, if we randomly choose one of these graphs, we *may* occasionally end up with a graph whose largest connected component has size $< \Theta N$. But with probability arbitrarily close to 1 as N is sufficiently large, we will draw a graph with a connected component of size $\geq \Theta N$. For the class of graphs $G_{ER}(N,\lambda)$ with fixed $\lambda > 1$, any value of Θ with $0 < \Theta < \varrho(\lambda)$ will work.

Here $\varrho = \varrho(\lambda)$ is a positive constant strictly between 0 and 1 that depends only on λ. It is the unique solution of the equation

$$1 - \varrho = e^{-\lambda\varrho} \tag{3}$$

in the interval $(0,1)$. It can be shown that for $\lambda > 1$ there exists exactly one such solution, while for $\lambda \leq 1$, no solutions of (3) fall in this interval. Let us list a few representative values of $\varrho(\lambda)$:

$$\begin{aligned} \varrho(1.1) &= 0.1761, \qquad \varrho(1.2) = 0.3137, \qquad \varrho(1.5) = 0.5828, \\ \varrho(1.75) &= 0.7127, \qquad \varrho(2.0) = 0.7968, \qquad \varrho(3.0) = 0.9405. \end{aligned} \tag{4}$$

One can prove that $\varrho(\lambda)$ is a strictly increasing function such that

$$\lim_{\lambda\to 1^+}\varrho(\lambda) = 0 \qquad \text{and} \qquad \lim_{\lambda\to\infty}\varrho(\lambda) = 1. \tag{5}$$

In other words, the expected relative size of the giant component increases with λ from a tiny fraction of all nodes to a fraction that comprises nearly all nodes.

For more information on the connected components of $G_{ER}(N,\lambda)$ we recommend the survey article [4].

Exercise 4 *(a) Are your results in Exercise 3 reasonably consistent with the theoretical predictions and the value* $\varrho(1.5) = 0.5828$*?*

(b) Try to confirm the theoretical predictions above by running a batch of 100 simulations each for instances of $G_{ER}(300, 0.5)$ *and* $G_{ER}(300, 2)$*.*

3 Disease transmission on Erdős-Rényi random graphs and complete graphs

The alert reader will have noticed the analogy between small *vs.* large components of $G_{ER}(N, \lambda)$ and minor *vs.* major outbreaks of diseases. There is in fact a close connection and we will examine it in this section.

Consider a next-generation compartment-based SIR-model with probability b of at least one effective contact until the next time step. This can be interpreted as a network-based model on a complete graph K_N. Consider a pair (i, j) of distinct hosts. Suppose host i becomes infectious at some time step t and host j will still be susceptible at time t. The probability that these hosts will have an effective contact during the interval of infectiousness of host i, that is, until step $t + 1$ in terms of the model, must be equal to b. The same will be true if host j becomes infectious at time step t and host i is still infectious at time t.

Now construct a random subgraph G of K_N as follows: Consider a simulation of a model as described in the previous paragraph that starts with one index case j^* in an otherwise susceptible population. One can think of the simulation as being performed by tossing biased coins sequentially as the simulation progresses, for all edges that have one endpoint that corresponds to a host who is infectious at the current time step, and whose other endpoint represents a host who is still susceptible at the current time step. The coin will need to be biased in such a way that it comes up heads with probability b. In terms of the simulation, the hosts that will be infectious at the next time step are exactly the ones represented by endpoints of those edges for which a coin is tossed and comes up heads. Include these edges in $E(G)$, and don't include those edges for which the coin comes up tails.

Exercise 5 *(a) If you have read Section 3 of [3], convince yourself that the description of the simulation given in the preceding paragraph matches the description of simulations given in that source.*

(b) Convince yourself that each edge of K_N *will be considered at most once in this construction of* $E(G)$ *for any given simulation, but there may be edges for which no decision about inclusion in* $E(G)$ *is made.*

In view of the result of Exercise 5(b), after the simulation has run its course, we still need to determine membership in $E(G)$ for those edges of K_N that have not yet been considered. Toss a coin that come up heads with probability b for each edge such edge and include it in $E(G)$ if, and only if, the coin comes up heads.

Exercise 6 *(a) Show that this construction will give an instance $G = G_{ER}(N, \lambda)$, where $\lambda = b(N-1)$.*

(b) What can you say about the size of the connected component of the index case in the resulting graph G? How is it related to the final size of the outbreak?

Notice that since we are considering a next-generation model, we are implicitly assuming that $\frac{\langle \tau^I \rangle}{\Delta t} = 1$. If N is large and b small, then $b(N-1) \approx bN$, and the formula for R_0 for the case of complete graphs K_N given in our review of models of disease transmission on networks at this web site[3] tells us that $R_0 \approx \lambda$, where λ is as in Exercise 6(a). Thus the theoretical predictions about the sizes of the connected components of $G_{ER}(N, \lambda)$ that we discussed in Section 2 imply the following results for network-based next-generation SIR-models with the uniform mixing assumption:

- If $R_0 < 1$, then a.a.s. all outbreaks will be minor.
- If $R_0 > 1$, then the final size $r(\infty)$ of major outbreaks will a.a.s. approach $\varrho(\lambda)$.
- If $R_0 > 1$, then the probability of $z(\infty)$ of minor outbreaks will a.a.s. approach $1-\varrho(\lambda)$.

You may want to compare these predictions with the the results we sketched in Subsection 2.3 of our review of models of disease transmission on networks at this web site. Note that here we get precise values for $r(\infty)$ and z_∞.

It may not be immediately obvious why the value of $z(\infty)$ will a.a.s. approach $1 - \varrho(\lambda)$ when $\lambda > 1$. To see this, consider a large batch of simulations for a fixed population size N and mean degree λ and construct one instance of $G_{ER}(N, \lambda)$ from each of these simulations. We could think about reversing the order of operations by first constructing instances of $G_{ER}(N, \lambda)$ as described in Subsection 1.1, then randomly drawing j^*, and then basing the simulations exactly on the same coin tosses that were used for the constructions of the random graphs. For large N, the probability that j^* does *not* end up in the largest component would be very close to $1 - \varrho(\lambda)$, and it would also be close to the probability of observing a minor outbreak in the given simulation.

These arguments explain why in simulations of next-generation SIR-models on Erdős-Rényi random networks we will observe similar outcomes in terms of the final size and the probability of minor outbreaks as in simulations on networks K_N with the same value of R_0. Such simulations were suggested in Module 6 of [2], and for them we observed that for *continuous-time models* the estimates for z_∞ are significantly lower for Erdős-Rényi random networks than for complete graphs K_N that embody the uniform mixing assumption. This discrepancy calls for an explanation, and we will give it next.

In continuous-time models, the duration of infectiousness τ_i^I of each node i is an exponentially distributed random variable. Most hosts will stay infectious for a period of less than $\langle \tau^I \rangle$, while some hosts will stay infectious for much longer than $\langle \tau^I \rangle$ (compare with Exercise 9.77 of Module 5 of [2]). These latter hosts will then on average cause many

[3] `http://www.ohio.edu/people/just/IONTW`

more secondary infections than R_0 would predict. This phenomenon does not occur in next-generation SIR-models where all hosts stay infectious for exactly one time step. Index cases with very long durations of infectiousness are much more likely to cause major outbreaks than index cases whose duration of infectiousness is shorter than $\langle \tau^I \rangle$. Since the nodes i with τ_i^I form a majority, we should expect z_∞ to be *larger* in continuous-time models than in next-generation models with the same R_0 on the same network. The effect is much more pronounced in complete networks where an index case who stays infectious for a very long time could in theory cause up to $N-1$ secondary infections than in Erdős-Rényi random networks where the maximum degree is expected not to exceed $\ln(N)$. This explains the discrepancies in the estimated values of z_∞ between next-generation and continuous-time models that were observed in the batch processing experiments of Module 6 of [2].

One can construct a graph G based on a continuous-time simulation in pretty much the same way as in the construction that we described above for simulations of next-generation models. The resulting graph G will be a random graph, but no longer an Erdős-Rényi random graph. To see this, consider two unordered pairs $\{j^*, i_1\}, \{j^*, i_2\}$, where i_1, i_2, j^* are all distinct. Each of these pairs will be included in $E(G)$ with probability b, but the events that $\{j^*, i_1\} \in E(G)$ and $\{j^*, i_2\} \in E(G)$ are no longer independent: If $\tau_{j^*}^I < \langle \tau^I \rangle$, each of these events will occur with probability less than b; if $\tau_{j^*}^I > \langle \tau^I \rangle$, each of these events will occur with probability larger than b. In other words, these two events will now be positively correlated. In contrast, in the construction of Erdős-Rényi random graphs, the decisions about which edges to include were all independent. Thus the probability distribution from which the random graphs G that we construct from continuous-time simulations are randomly drawn will be different from the distribution of Erdős-Rényi random networks $G_{ER}(N, \lambda)$.

References

[1] P Erdős and A Rényi. On the evolution of random graphs. *Selected Papers of Alfréd Rényi, vol*, 2:482–525, 1976.

[2] Winfried Just, Hannah Callender, and M Drew LaMar. Disease transmission dynamics on networks: Network structure *vs.* disease dynamics. In Raina Robeva, editor, *Algebraic and Discrete Mathematical Methods for Modern Biology*, pages 217–235. Academic Press, 2015.

[3] Winfried Just, Hannah Callender, M Drew LaMar, and Natalia Toporikova. Transmission of infectious diseases: Data, models, and simulations. In Raina Robeva, editor, *Algebraic and Discrete Mathematical Methods for Modern Biology*, pages 193–215. Academic Press, 2015.

[4] Joel Spencer. The giant component: The golden anniversary. *Notices of the AMS*, 57(6):720–724, 2010.

Sample solutions for the exercises

Sample solution for Exercise 1: In some runs, almost no secondary infections occur, which indicates that the connected component of j^* is very small. In other runs, more than half of all nodes get infected. This indicates that j^* belongs to a large component that comprises more than half (for a typical instance, around 58%) of all nodes. It must be the same component each time, since different connected components must be disjoint and there cannot be 2 distinct ones that contain more than half of all nodes. □

Sample solution for Exercise 2: The results are similar to those of the previous exercise in that we observe a the same dichotomy between very small connected components and one very large component. However, the relative size of the large component now fluctuates around 0.58. The reason is that we create a new instant of $G_{ER}(300, 1.5)$ for each run. □

Sample solution for Exercise 3: In our experiment we observed a distinct gap between 34 runs where j^* belonged to a component of size at most 10 (a proportion of at most 0.03 of all nodes) and 66 runs where j^* belonged to a component of size at least 115 (a proportion of at least 0.38 of all nodes). The mean size of these large components was 174.1, which represents a proportion of 0.58 of all nodes. These results confirmed our observation in the previous exercises. □

Sample solution for Exercise 4: (a) Yes, the results of Exercise 3 were remarkably close to the theoretical prediction.

(b) In our experiment with 100 runs for $G_{ER}(300, 2)$ we observed 20 runs where j^* belonged to a component of size at most 4 (a proportion of at most 0.013 of all nodes) and 80 runs where j^* belonged to a component of size at least 206 (a proportion of at least 0.69 of all nodes). The mean size of these large components was 239.15, which represents a proportion of 0.7972 of all nodes.

In our experiment with 100 runs for $G_{ER}(300, 0.5)$ we did not observe a distinct gap between small and large components. In 64 runs, j^* was an isolated node. The largest connected component of j^* that we found had size 10, which represents a proportion of 0.033 of all nodes.

These results were remarkably close to the theoretical predictions. □

Sample solution for Exercise 5: (b) A given edge $\{i, j\}$ will be considered only at the first time step t when either node i or node j is infectious. Since we are assuming an SIR-model, after this time step, at least one of the nodes i, j will be removed. If neither i nor j become infectious during the outbreak or if both of these nodes become infectious at the exact same time, no decision about inclusion of the edge $\{i, j\}$ will be made. □

Sample solution for Exercise 6: (a) Will be given in our module *Mathematical models and theorems.*

(b) The connected component of j^* is the subgraph induced by all nodes that experience infection. Its relative size is the final size of the simulated outbreak. □