Exploring distances with IONTW*

Winfried Just† Hannah Callender‡ M. Drew LaMar§

December 23, 2015

This module has two parts. The first part is purely conceptual and invites readers to critically evaluate popular claims based on Stanley Milgram's famous experiment that gave birth to the phrases "small-world property" and "six degrees of separation." In the second part we use IONTW to explore distances between nodes in several types of networks. We also explore various possible formal definitions of the small-world property and propose one that is suitable for classes of disconnected networks. The parts are somewhat independent of each other, but we recommend that the reader work through both parts in the order given here. The exercises within each part should definitely not be attempted out of sequence.

1 Milgram's famous "six degrees of separation" experiment

Consider the network G_{FN} whose nodes represent humans and whose edges connect any two persons who are acquainted on a first-name basis. Let $d(i,j)$ be the distance between two nodes in this network, that is, the number of edges in the shortest path from i to j in G_{FN}. If i,j are chosen randomly, then $d(i,j)$ becomes a r.v. What can we say about its distribution?

The American social psychologist Stanley Milgram and his collaborators conducted an ingenious experiment to answer this question; the results were reported in [2]. The experimenters recruited 296 volunteers from Nebraska and the Boston area. Each volunteer i was given a letter with some information about a Boston stock broker j and was instructed to send it to a person with whom the volunteer was acquainted on a first-name basis (along an edge of G_{FN}) and whom the volunteer thought to be at a closer distance from j in G_{FN}. Attached to the letter were instructions to continue forwarding it in this manner, until it would be sent to the stock broker j. The experimenters kept track of the number of intermediaries $i_1, \dots, i_m$ (excluding i and j) that forwarded the letter.

Exercise 1 *Suppose the letter did eventually get sent to j. How is the number m of intermediaries related to the distance $d(i,j)$?*

*©Winfried Just and Hannah Callender, 2014

†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

Out of the 296 letters, 64 eventually reached their target.

Exercise 2 *Does the success rate of $\frac{64}{296}$ (approximately 22%) tell us anything about the structure of the network G_{FN}?*

For the letters that did arrive, the researchers reported a mean number of 5.2 intermediaries. This result has inspired the popular claim that there are only *six degrees of separation* between any two humans.

Exercise 3 *Critically evaluate this claim. What do or don't the results of Milgram tell us about the likely maximum, mean, or median of $d(i,j)$ in G_{FN}?*

We will return to this exercise at the end of the second part of this module, but don't try to look up our solution at this point. Instead, write down your thoughts.

2 Exploring distances with IONTW

Open IONTW, click **Defaults**, and choose the following parameter settings:

network-type → **Nearest-neighbor 2**
num-nodes: 100
d: 1

Create a network by pressing **New** and then toggle **Labels** to see how the nodes are numbered. The network that you see in your **World** window will be denoted by $G^2_{NN}(100,1)$. We call the graphs $G^2_{NN}(N,d)$ *two-dimensional nearest neighbor networks.*

Exercise 4 *(a) Find $d(13,21)$ and $d(21,66)$ in this network.*

(b) Find the diameter of this network, that is, find the maximum distance of any pair of nodes. For which pairs of nodes is the maximum attained?

Calculating the mean or median distances for randomly chosen nodes in $G^2_{NN}(100,1)$ would be a bit tedious. Can you look it up in the **Command Center** after pressing **Metrics**?

Well, maybe. The display gives you a line

```
Average path length in largest component = 6.66667
```

This looks like it could be the mean or median distance. But the word "Average" could refer to a mean or a median, and the phrase "path length" could refer to distance (the length of the shortest path) or to some kind of average taken over all paths. We could have been more clear, but we designed IONTW so as to put you here in a situation that you will encounter quite frequently when using commercial software: The terminology in the interface is not what you are used to, and surprisingly often it is ambiguous. When using software for scientific purposes, you always need to make sure what, exactly, it actually calculates. Let us divulge that this metric is indeed a mean or median distance. But this leaves you still with as many as 4 distinct interpretations, depending on whether or not distances from nodes to themselves are included in the calculations!

Let us see which interpretation is correct. Change

num-nodes: 6

and click **New** to create a network $G^2_{NN}(6,1)$. Then click **Metrics** to see which value IONTW reports for

```
Average path length in largest component
```

Exercise 5 *(a) Find the median distance for all 15 pairs $\{i,j\}$ of nodes with $i \neq j$.*

(b) Find the mean distance for all 15 pairs $\{i,j\}$ of nodes with $i \neq j$.

(c) Find the median distance for all 21 pairs $\{i,j\}$ of nodes.

(d) Find the mean distance for all 21 pairs $\{i,j\}$ of nodes.

*(e) Which of these results agrees with the value that is reported in the **Command Center** after pressing **Metrics**?*

Let us look at another example of how the mean distance can be related to the median distance. Choose the following parameter settings:

network-type → **Regular Tree**
lambda: 2
d: 5

Create a network by clicking **New**, move the speed control slider to the extreme right, press **Spring**, and wait a minute or so. When the graph in the **World** window has taken a nice shape, press **Spring** again and then **Scale** to make it better fit the **World** window. One needs to be a little careful with interpreting the resulting picture in the **World** window. All nodes numbered 6 or higher are leaves of this tree and have degree 1, even though the action of **Spring** may place some of them on top of another edge. Choose

plot-metric → **Shortest Paths**

and click **Update**. The height of the largest bar of the **Network Metrics** plot is shown on the vertical axis; you can find the approximate height of the other bars in the histogram by moving the cursor to the top line of each bar.

Exercise 6 *Use the histogram in the **Network Metrics** plot to find the median distance for this network and compare it with the mean.*

The mean distance and the diameter are properties of a given network. Let us see what they tell us about the spread of diseases. Click **Defaults** and choose the following settings:

model-time: Discrete
time-step: 1
infection-prob: 1
end-infection-prob: 1

network-type → **Nearest-neighbor 2**
num-nodes: 100
d: 1
set-state-by → **Vector from input**

This sets up a next-generation SIR-model where at each time step every host is guaranteed to make effective contact with all adjacent hosts.

Create a **New** network $G^2_{NN}(100, 1)$, and toggle **Labels** to see the numbering of the nodes. Make node 0 the index case by using the following procedure:

Click **Set**
In the dialogue box that appears enter
`[0]`
Click **OK**.

Make sure that everything worked as expected so that you have an initial state where node 0 is infectious and all other nodes are susceptible. Set the speed slider to a slow speed; adjust for comfortable viewing as needed. Now click **Go** and watch the movie. If you want to restart it, use **Last.**

Exercise 7 *(a) What is the relationship between node 0 and the nodes that are infectious at time step t?*

(b) How is the time at the end of the outbreak, measured in `ticks`*, related to the network properties that we discussed earlier?*

Now repeat the experiment by using **Reset** and then making first node 50 and then node 55 the single index case. In the dialogue box that appears after clicking **Set** you will need to enter first `[50]` and then `[55]`.

While previously the duration of the outbreak was $diam(G^2_{NN}(100, 1)) + 1$, now you get shorter outbreaks. But even after subtracting 1, you will get a number that exceeds the mean distance. Let us use the symbol $D(j^*)$ for the number that we get after subtracting 1 from the duration of the outbreak caused by index case j^* in an otherwise susceptible population in a model with the disease transmission parameters listed above. This number can be defined for every network.

Exercise 8 *(a) Give a verbal description of $D(j^*)$ and convince yourself that it cannot exceed the diameter of the network.*

(b) Convince yourself that in any network the mean $\langle D(j^*)\rangle$ cannot be smaller than the mean distance between nodes.

(c) Can you find networks where the mean $\langle D(j^*)\rangle$ is equal to the mean distance between nodes?

We have observed something very important: Two *network parameters,* the diameter and the mean distance, give us upper and lower bounds on the mean duration of outbreaks of certain types of diseases. Thus one could use these network parameters to make predictions about the duration of outbreaks. The relation between these network parameters and the duration of outbreaks becomes less tight for disease transmission parameters that are different from the ones we considered here, but there will still be a close connection.

Unfortunately, in simulations we are restricted to exploring rather small networks, while real outbreaks happen in populations of thousands or millions of hosts. It is therefore of interest to explore how the diameter and mean distance scale if we retain the general structure of the network but increase the number of nodes.

Let us start with very simple networks. Change the following settings:

Press `Clear` on the **Command Center** bar.
network-type → **Nearest-neighbor 2**
num-nodes: 11
d: 1

Click **New** and find the diameter of the network by visual inspection. Record it, and then click **Metrics** to verify your results. Repeat with

num-nodes: 23, 47

Note that for prime numbers N the graphs $G^2_{NN}(N, 1)$ are not really two-dimensional grids but simple paths.

Enlarge the **Command Center** with the double-arrow icon and look at the data that you have recorded. Would it be fair to say that for graphs $G^2_{NN}(N, 1)$ with N prime the diameter and the mean distance roughly double when you roughly double the number of nodes?

Such a pattern is indicative of *linear scaling.* We say that the value of a quantity $\chi(N)$ that depends on N *scales linearly* if

$$\lim_{N\to\infty} \frac{\chi(N)}{N} = c, \tag{1}$$

where c is a nonzero constant. This does not mean that $\chi(N) = cN$, it only means that for sufficiently large N we can take cN as a fairly good estimate of $\chi(N)$. Since $c(2N) = 2cN$, linear scaling produces the pattern that you observed.

Nonlinear scaling can take many different forms. For example, if D is a fixed exponent, then we say that $\chi(N)$ *scales like* N^D if

$$\lim_{N\to\infty} \frac{\chi(N)}{N^D} = c \neq 0. \tag{2}$$

Of course, (1) is nothing else than the special case of (2) for $D = 1$. In other words, linear scaling is a special kind of *power law scaling.* Now suppose $\chi(N)$ scales like $N^{0.5} = \sqrt{N}$. Then quadrupling N should have the effect of roughly doubling $\chi(N)$ as $c\sqrt{4N} = \sqrt{4}cN$. Let us see whether the class of square grids without diagonals is such an example.

Repeat the previous experiment for

num-nodes: 25, 100, 400

Does it appear the the diameter and mean distance scale like $\sqrt{N}$ in the class of networks $G^2_{NN}(N,1)$ with $N = n^2$?

So far, we have considered only networks with a rigid structure. But now let us explore some random graphs. Change the following parameter settings:

Press `Clear` on the **Command Center** bar.
network-type → **Random Regular**
num-nodes: 50
lambda: 4

Click **New** to create an instance of $G_{Reg}(50,4)$. Use **Metrics** to find the mean distance between nodes.

Since every instance of a random graph is slightly different, we may want to look at several instances drawn from the same distribution. Create 5 instances each of $G_{Reg}(50,4)$, $G_{Reg}(100,4)$, $G_{Reg}(200,4)$ by changing **num-nodes** accordingly and using **New**. For each new instance, click **Metrics**.

Enlarge the **Command Center** with the double-arrow icon and look at the data that you have recorded. Would it be fair to say that the mean distance between nodes roughly increases by a constant number when you double the number of nodes?

This pattern is indicative of *logarithmic scaling* rather than power law scaling. We say that $\chi(N)$ *scales logarithmically* if

$$\lim_{N\to\infty} \frac{\chi(N)}{\ln(N)} = c \neq 0. \tag{3}$$

Since $c\ln(2N) = c(\ln(2)+\ln(N)) = c\ln(N)+c\ln(2)$, if $\chi(N)$ scales logarithmically, then for sufficiently large N we would see a roughly constant increment of $\chi(N)$ by $\approx c\ln(2)$ when we double N.

Recall our discussion on Section 1 of the famous experiment that gave birth to the phrase "small-world property." If you have initially skipped this section, work through it now. Why would a mean or even maximum distance of 6 between two randomly chosen people be considered surprisingly small? There would not be anything remarkably small about this number in a village of 100 people or even in the town of 20,000. But on the scale of the whole human population, the number 6 seems surprisingly small. We can see that a rigorous definition of the *small-world property* will make sense only in the context of a class of networks, and it should be expressed in terms of a scaling law of an average distance. As a first approximation, let us say that a class of networks has the small-world property if the average distance for networks in this class scales *at most* logarithmically with network size N. This means that (3) will hold if we take $\chi(N)$ to be the average distance and allow the limit to be zero. Equivalently, we could say that for some fixed $c > 0$ we will have $\chi(N) \leq c\ln(N)$.

We have again been vague about the meaning of "average" in the above definition. For classes of connected graphs we could use the mean. The classes $G^1_{NN}(N,d)$ and $G^2_{NN}(N,d)$ of nearest neighbor networks consist of connected graphs and our explorations indicate that they do not have the small-world property. This can be rigorously proved. Random regular graphs $G_{Reg}(N,k)$ for $k > 2$ are a.a.s. (asymptotically almost surely[1]) connected, and it can be rigorously proved that for any $k > 2$ this class has the small-world property. Our explorations indicate as much. But since, for example, $G^1_{NN}(200,2)$ is also an instance of a 4-regular graph of size N, it *could* be drawn as an instance of $G_{Reg}(200,4)$, although the probability of this event is very, very small. For classes of random graphs we will in general not be able to say that the average distance is *always* $\leq c\ln(N)$. This inequality will only hold a.a.s., that is, with probability arbitrarily close to 1 as $N \to \infty$.

But how about disconnected graphs? Defining the small-world property in terms of the mean distance no longer works as for a graph with several connected components. As we explained in our brief review of probability theory at this web site, the mean distance will always be infinite. Should we interpret the "average" as the median in this case?

Consider the class of Erdős-Rényi random graphs $G_{ER}(N,\lambda)$ for a fixed $\lambda > 1$. According to the results of our module *Exploring Erdős-Rényi random graphs with IONTW* at this web site, the graphs in this class are a.a.s. disconnected, with one giant connected component of size close to $\varrho(\lambda)N$, and the size of the second largest component scaling logarithmically in N. Thus a.a.s., the mean distance between two randomly chosen nodes will be infinite. However, the mean distance between any two nodes in the *largest component* will still be finite, and this is the value that our software reports when you press **Metrics** after creating a **New** network. The corresponding median is not reported; we need to estimate it ourselves. The diameter of the largest connected component will be helpful in these estimates. One can prove that this diameter a.a.s. scales logarithmically for graphs $G_{ER}(N,\lambda)$ [1].

Exercise 9 *How could you use IONTW to obtain a rough empirical confirmation of this theoretical result by considering instances of $G_{ER}(N,4)$ for $N = 50, 100, 200$?*

In other words, for fixed $\lambda > 1$, a.a.s. the diameter of the largest connected component of $G_{ER}(N,\lambda)$ will be $\leq c_{diam}\ln(N)$ for some fixed constant c_{diam}. If the inequality holds, then $d(i,j) \leq c_{diam}\ln(N)$ for all pairs of nodes $i \neq j$ that both belong to the giant component. But if i and j belong to *different* connected components, then $d(i,j) = \infty$. Thus the overall median of the distances will be $\leq c_{diam}\ln(N)$ if the probability that two randomly chosen distinct nodes both belong to the giant component is at least 0.5, and will be infinite otherwise.

Now let us return to the problem of defining the small-world property for classes of disconnected graphs. We could perhaps say that the class of Erdős-Rényi random graphs with mean degree λ has the small-world property if a.a.s. the *overall* median distance does not exceed the diameter of the giant component. But when will this be the case?

Let us look at some examples. Change the following parameter settings:

network-type → Erdos-Renyi

[1] See Subsection 1.2 of our module *Exploring Erdős-Rényi random graphs with IONTW* at this web site `http://www.ohio.edu/people/just/IONTW` for a definition of "a.a.s."

num-nodes: 200
lambda: 2

Click **New** to create an instance of $G_{ER}(200, 2)$. According to the logarithmic scaling law for the diameter, we can assume that the largest components of the graph that you see in the **World** window will with high probability have diameter $\leq c_{diam} \ln(N)$. Now press **Metrics** and use the information it gives you about the largest component to complete the following exercise.

Exercise 10 *(a) Estimate the probability that two randomly chosen nodes belong to the largest component of the instance of $G_{ER}(200, 2)$ that you see in your **World** window. Based on this calculation, is the overall median distance finite or infinite?*

*(b) Now change **lambda** to 1.5 while retaining all other parameters. Create a **New** network and repeat the calculations of point (a) for the new network.*

(c) Formulate necessary and sufficient conditions on the parameter λ for the overall median distance Q_2 to a.a.s. satisfy the inequality $Q_2 \leq c_{diam} \ln(N)$.

Most sources in the literature define the small-world property in terms of logarithmic scaling of the mean distance. This works fine for connected graphs, but even a single isolated node makes the mean distance infinite. Exercise 10 shows that the median overall distance may or may not work well for classes of disconnected graphs. If you analyze its solution carefully, you will notice that for each $\lambda > 1$ there is always a positive P so that in the class of random graphs $G_{ER}(N, \lambda)$ the P-th percentile of the distance between randomly chosen nodes scales at most logarithmically: Any P with $0 < \frac{P}{100} < (\varrho(\lambda))^2$ will work.

Thus we believe that the best way of conceptualizing the small-world property of a class of graphs is to require that for some fixed $P > 0$ there exists a positive constant c such that the P-th percentile of the distance between randomly chosen nodes a.a.s. satisfies the inequality $\leq c \ln(N)$. The classes of random regular graphs $G_{Reg}(N, k)$ have this property for all choices of $k > 2$, but do not have it for $k = 1$ or $k = 2$. The classes of Erdős-Rényi random graphs $G_{ER}(N, \lambda)$ have this property for all choices of $\lambda > 1$, but do not have it for $\lambda \leq 1$. The classes of graphs $G^1_{NN}(N, d)$ or $G^2_{NN}(N, d)$ do not have this property for any choice of d.

Exercise 11 *Reread your solution of Exercise 3. How do your arguments relate to what you have learned in this second part of the module? Would you want to modify your solution?*

References

[1] Bela Bollobás. *Random graphs.* Cambridge University Press, second edition, 2001.

[2] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.