Random Graphs and Critical Connectivity

Michael Hughes

January 15, 2009

A popular area of study within graph theory considers the properties of *random* graphs. There are many different ways one could think of to construct a graph probabilistically. One of the most simple and most fascinated was first documented by mathematicians Paul Erdos and Alfred Renyi in the 1950s. The Erdos-Renyi random graph model assigns each possible edge in a n-vertex undirected graph a fixed probability p of existence. Each edge's existence is determined independently of its neighbors, though every edge uses the same value of p.

With this model, we can tackle a very interesting experiment. First described by Erdos and Renyi in 1960, the experiment tests the connectivity of random Erdos-Renyi graphs over many values of p. Recall that in a connected graph, there exists a path between all possible pairs of vertices.

The fascinating conclusion of this experiment observes that there exists some critical value p^* such that for $p < p^*$ almost all graphs are disconnected and for $p > p^*$ almost all graphs are connected. This phenomena is known as a phase change or *percolation* due to the abrupt change in the connectivity property for a small change in the parameter p. In addition to this percolation, the trend between graph size n and the critical point value p^* is also notable, and will form the focus of the rest of this chapter.

We'll tackle this problem in a few ways. First, we implement a simulation to observe the percolation phenomenon firsthand and gain some sense of its dependence on n. Next, we use theoretical analysis to discover a closed-form expression relating graph size to critical point value. Finally, we validate our analytical results with more simulation.

1 Simulation Study of Critical Point

To begin, we create a simulation for studying critical points. The idea behind the simulation is simple: for a given n value, we iterate over a linear sequence of p values and for each p we construct many random graphs. We test the conductivity of each of these graphs and record the fraction of tested graphs which are connected.

Using the `pylab` module of Python, we can visualize this data and observe

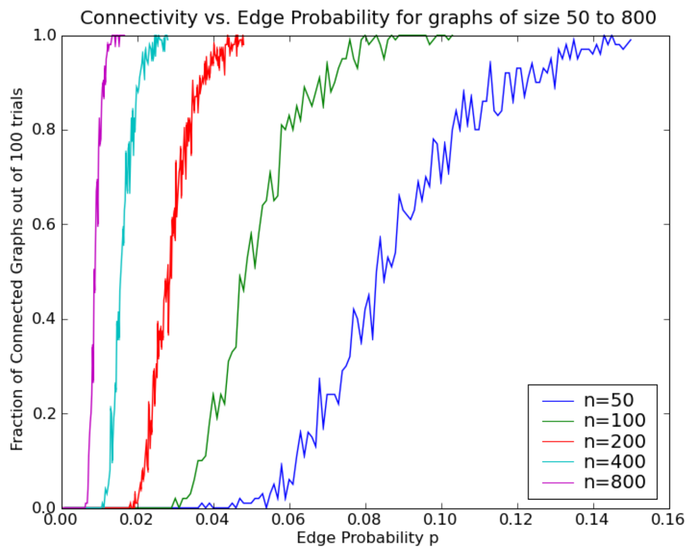

Figure 1: Fraction of connected graphs as a function of edge probability p

the percolation phenomena firsthand. A graph of observed connectivity probability versus edge probability appears in Figure 1.

First, we notice that our simulation appears to be working, because for each n series we notice that graphs are very unlikely to be connected at low p and highly likely to be connected for high p. Furthermore, we can observe trends in the critical point as n increases. It appears that larger graph sizes have more abrupt phase transitions. That is, the window of p values over which the percolation effect happens becomes narrower as n gets larger. Additionally, we see that the actual value of p^* appears to decrease as n increases. The space between critical points decreases while n increases by the same factor. No asymptotic limit to p^*'s movement is visible, so it may be possible that it will decay toward zero as $n \to \infty$.

1.1 Theoretical Observations on the Critical Point

We now think analytically about random graphs and connectivity. Specifically, let's figure out exactly how the critical point p^* depends on n. Surprisingly, it is not too difficult to establish the exact asymptotic value of the critical point as a function of n. In fact, in this section we show that, as n approaches ∞, the critical point is $p^* = \frac{\ln(n)}{n}$.

This work is heavily indebted to Gopal Pandurangan, a computer science professor at Purdue, who published an online monograph on this subject. The document is available at `http://www.cs.purdue.edu/homes/gopal/CS590A-2007/11.pdf`.

To prove this result, we first must reason carefully about what it means to be connected in general. We observe that for any graph to be connected, it must have no disjoint components. In other words for graph G, if we find a non-empty subset of vertices A such that no vertex in A has an edge connecting it to vertices outside of A, then G is unconnected.

With this notion in mind, we now consider what this means probabilistically speaking. That is, given only the probability information about the graph's edges, how can we say conclusively whether or not some edges exist, much less that it is or is not connected? The answer to this conundrum is to borrow the concept of *expected value* from probability theory. We'll assume the reader has a cursory knowledge of this concept (for a brief primer, check out the Wikipedia article: `http://en.wikipedia.org/wiki/Expected_value`).

Our argumentation consists of two parts. First, we show that as a graph gets arbitrarily large, the most common disjoint, isolated component A is the single vertex ($|A| = 1$). That is, as $n \to infty$, the expected number of isolated single vertices is always more than the expected number of disjoint subcomponents of size 2 or 3 or 4 or ... however big we like (while $a < n$ of course). Secondly, we show that as graph size increases, the edge probability p has a critical point $p^* = \frac{\ln(n)}{n}$ such that for $p < p^*$ the expected number of isolated single vertices gets arbitrarily large and graphs are almost always disconnected, while for $p > p^*$ the expected number of isolated single vertices approaches zero (squashing the expected number of larger disjoint subcomponents to zero in the process) and makes these graphs almost always connected.

Now, to the proof. First, given n and p, we can come up with an expression for the number of expected isolated components of a particular size a, where $a < n$. Let's first think about the easy situation when $a = 1$, which corresponds to a single isolated vertex v (no edges connect to it). There are totally $n - 1$ possible edges v could have had, one for each other vertex in the graph G. The probability that an edge between two nodes does NOT form is $(1-p)$. So we can easily see that the probability that a particular vertex is isolated (unconnected) is

$$P_{isolated,|A|=1} = (1-p)^{n-1}$$

Given that each node in the graph has the same probability to be isolated, we can find the expected value of isolated nodes (in other words, the number of nodes on average which will be isolated in a (n, p) graph) as follows

$$E_{isolated,|A|=1} = \sum_{i=1}^{n} (1-p)^{n-1} = n(1-p)^{n-1}$$

These results can quickly be extended to disjoint, multi-node subcomponents of the graph. For such a component A containing a nodes, we find that each node v_a in A has $n - a$ possible nodes outside of A it could form an edge with. Thus, the probability that a particular vertex v_a did not form any of those edges is $(1 - p)^{n-a}$. Repeating this for all a nodes gives the probability that every node in A is isolated with respect to $\overline{A}$ as

$$P_{isolated,|A|=a} = (1 - p)^{a(n-a)}$$

There are $\binom{n}{a}$ possible subcomponents of size a in a given graph. So the expected number of disconnected subcomponents with size a is

$$E_{isolated,|A|=a} = \sum_{i=1}^{\binom{n}{a}} (1 - p)^{a(n-a)} = \binom{n}{a}(1 - p)^{a(n-a)}$$

We now have a firm expression for how many components of a given size a we can expect to be disconnected with in a random graph with size n and edge probability p. So how do we actually pin down the critical point p^*? Remember, graphs with p such that $p < p^*$ will almost surely be disconnected while $p > p^*$ graphs will almost surely be connected.

Our next step is to examine how connectivity changes as graph size gets bigger. We'll take the typical mathematical approach and examine how the number of expected isolated components changes as $n \to \infty$.

Let's determine what size of isolated components will be most likely to occur as graph size increases. In other words, for some p, as we let n grow will isolated single vertices be the most common disconnected component? Or will we be more likely to find a larger set of vertices (say a group of 2 or 3 or a nodes) randomly isolated from the graph?

One way to answer this question is to determine whether our expectation expression is monotonic with a as $n \to \infty$. In other words, is it always more likely to find isolated component with size a than it is to find $a + 1$? We'll hypothesize that this is true, and then prove it. To start, we have

$$\begin{aligned} E_{isolated,|A|=a} &\overset{?}{>} E_{isolated,|A|=a+1} \\ \binom{n}{a}(1 - p)^{a(n-a)} &\overset{?}{>} \binom{n}{a+1}(1 - p)^{(a+1)(n-a-1)} \end{aligned}$$

Through basic algebraic manipulation, we can work these terms as follows

$$
\begin{aligned}
\binom{n}{a}(1-p)^{a(n-a)} &\overset{?}{>} \binom{n}{a+1}(1-p)^{(a+1)(n-a-1)} \\
\frac{n!}{a!(n-a)!}(1-p)^{a(n-a)} &\overset{?}{>} \frac{n!}{(a+1)!(n-a-1)!}(1-p)^{(a+1)(n-a-1)} \\
\frac{(a+1)!(n-a-1)!}{a!(n-a)!} &\overset{?}{>} \frac{(1-p)^{(a+1)(n-a-1)}}{(1-p)^{a(n-a)}} \\
\frac{(a+1)}{(n-a)} &\overset{?}{>} \frac{(1-p)^{(a+1)(n-a-1)}}{(1-p)^{a(n-a)}}
\end{aligned}
$$

Letting $n \to \infty$, we know that because $n > a$ we can simplify $n - a \approx n$

$$
\begin{aligned}
\frac{(a+1)}{n} &\overset{?}{>} \frac{(1-p)^{(a+1)(n)}}{(1-p)^{a(n)}} \\
\frac{(a+1)}{n} &\overset{?}{>} (1-p)^{(a+1)n-an} \\
\frac{(a+1)}{n} &\overset{?}{>} (1-p)^{n}
\end{aligned}
$$

Pushing all n terms to the right side, we can see that in the limit $n \to \infty$, the $(1-p)^n$ term will go to zero much faster than the n term will approach infinity, so

$$
a+1 \overset{?}{>} 0
$$

We've got it! We've now shown that as $n \to \infty$

$$
E_{isolated,|A|=a} > E_{isolated,|A|=a+1}
$$

So the expected number of isolated components is monotonically strictly decreasing as $n \to \infty$. Thus, the most common isolated component in a random n, p graph as graph size increases will be the lone isolated vertex (corresponding to $a = 1$).

With this knowledge, we can now restrict our connectivity analysis to single isolated vertices and move to the second major portion of our analysis. Our new question is: as graph size increases, what value of p will balance critically between when the expected number of isolated vertices grows to infinity (making the graph disconnected) or drops to zero (a connected graph)?

Let's return to our expression for expected lone vertices.

$$E_{isolated,|A|=1} = n(1-p)^{n-1}$$

We can rearrange this to get a better idea of how p might influence the expected number E. We can rewrite this as an exponential equation

$$E_{isolated,|A|=1} = ne^{(n-1)\ln(1-p)}$$

Remembering Taylor approximations, we recall that

$$\ln x = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}(1-x)$$

So we can rewrite the $\ln(1-p)$ term as

$$\ln(1-p) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}(p) = -p + -\frac{p^2}{2} + -\frac{p^3}{3} + \cdots$$

Because p is a probability, we know that $0 < p < 1$. Thus $|p| >> |p^2| >> |p^3| >> ...$, and we make the approximation

$$\ln(1-p) \approx -p$$

Returning to our expectation calculation, our expectation becomes

$$E_{isolated,|A|=1} \approx ne^{-p(n-1)}$$

We still don't know how E depends on p critically. So why don't we make an educated guess about how p depends on n asymptotically. For reasons that will hopefully become clear later, let's assign $p = c\frac{\ln n}{n}$, where c is some non-negative real constant. Our equation is now

$$E_{isolated,|A|=1} \approx ne^{-c\frac{\ln n}{n}(n-1)}$$

As $n \to \infty$, the n in the denominator cancels the $(n-1)$ term, and this becomes

$$\begin{aligned}
E_{isolated,|A|=1} &\approx ne^{-c\ln n} \\
E_{isolated,|A|=1} &\approx ne^{\ln(n^{-c})} \\
E_{isolated,|A|=1} &\approx n(n^{-c}) \\
E_{isolated,|A|=1} &\approx n^{1-c}
\end{aligned}$$

We can know understand how the expected number of isolated vertices varies with c. If $c < 1$, then n^{1-c} will increase as $n \to \infty$, while if $c > 1$, the exponent will be negative and the expression will decay toward zero. So we have the piecewise conditions

$$E_{isolated,|A|=1} \to \begin{cases} 0 & \text{if } p > \frac{\ln n}{n} \\ \infty & \text{if } p < \frac{\ln n}{n} \end{cases} \quad \text{as } n \to \infty$$

Or in practical terms

$$\text{Almost all graphs are} \begin{cases} \text{connected} & \text{if } p > \frac{\ln n}{n} \\ \text{disconnected} & \text{if } p < \frac{\ln n}{n} \end{cases}$$

Finally! We know that if we let $p = c\frac{\ln n}{n}$, then as graph size increases we know that the typical graph will be very likely connected if $c > 1$ and very likely disconnected if $c < 1$. Thus, we have a critical point $p^* = \frac{\ln n}{n}$.

1.2 Validation

To confirm that our expression for the critical point makes practical sense, we can estimate the critical point for various n values in simulation and compare the result to the expected theoretical value. Figure 1.2 shows such a plot for n values from 50 to 800 geometrically increasing by a factor of 2.

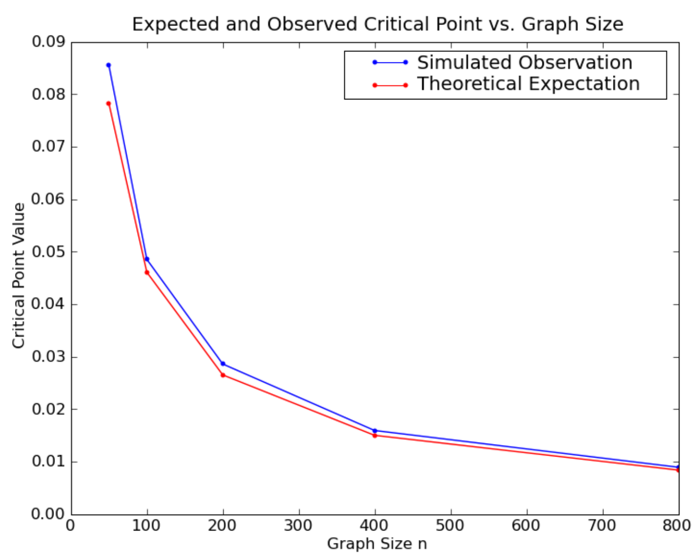

Figure 2: Simulated and Expected Critical Point Values for various graph sizes.

We see that our simulation confirms our theoretical expectation extremely well. There is very little difference between the two curves, indicating that our expression for the critical point seems to be correct.