Phase Transitions in Random Graphs-Outbreak of Epidemics to Network Robustness and fragility

Mayukh Nilay Khan

May 13, 2010

Abstract

Inspired by empirical studies researchers have tried to model various systems like human populations, the World Wide Web or electric power grids by random graphs. Here we first examine different properties of random graphs(both undirected and directed) having arbitrary degree distributions using the generating functon formalism. We present some empirical data about the structure of random graphs in real life especially the WWW. Then we modify our results to include site and bond percolation to address questions about the onset of large scale connectivity and the formation of a giant component in random graphs. This point corresponds to a phase transition. We use this information to examine questions about the onset of epidemic or the robustness or fragility of a network to random or targeted deletion of nodes.

1 Introduction -Motivation and some general considerations for the model

A random graph is a collection of points or vertices with lines or edges connecting them at random. Mathematicians have been studying random graphs for a long time, starting with the work by Paul ErDös and Alfrèd Rènyi. The problem of applying their results to real life networks is as follows:

If we assume each edge in the graph to be present with an independent probability p, for a graph with N vertices where each vertex is connected on an average to z edges, then $p = z/(N-1)$, which is approximately equal to z/N in the limit of large N. From here and now on p_k is used to denote the probability that a randomly chosen vertex on the graph has degree k, i.e. it is connected to k other vertices. Hence,

$$p_k = Lt_{N\to\infty}\binom{N}{k}p^k(1-p)^{N-k} \simeq \frac{z^k\exp(-z)}{k!} \tag{1.1}$$

However, real world networks (for e.g. the WWW, where each page is a vertex and the hyperlinks are edges, a directed graph) are not poissonian. In the specific case of the internet we see that the graphs have power law behaviour as explained later. Hence it is necessary to modify these results for graphs with arbitrary degree distributions and also on the nature of the graphs (directed or undirected). Then we generalize our results to include site and bond percolation on these graphs.

Now,let us understand why it is "useful" to study such networks. First we translate the language of random graphs to real world networks and then address questions about them. The electric power grids, airline networks and the Internet are examples of networks whose functionality depends on the pattern of interconnection between their nodes, so it is important to study such networks and their fragility to the removal of nodes. We study the question of connectivity by determining when a giant cluster with large scale connectivity forms.

Nodes on the graph are considered to be occupied if they(power grid or router) are functioning properly. We can consider occupation probability as a function of vertex degree. We observe percolation clusters on the graph as the occupation probability is varied. We see that if the structure of the network is chosen appropriately we can make the network resilient to random deletion of nodes.

If we want to check the connectivity of the network to failure of links(transmission cables, optic fibres) between the nodes we can study the problem of bond percolation on the network. For the problem of disease propagation, occupied vertices on the graph are people who are already infected or susceptible to the disease. Links represent contacts capable of spreading the disease and can be assumed to be occupied with some probability representing that only some contacts actually lead to transmission. In the language of percolation phase transition there is a point as a function of occupation probability when a giant component(with occupied vertices) forms. In the language of power networks this is when the graph achieves large scale connectivity,so it works efficiently as a distribution system, in the language of disease propagation this is the point when a large portion of the population is infected and this point corresponds to the outbreak of an epidemic. As, in the case of phase transitions in physical systems we will see that it makes sense to say a phase transition occurs only when the system size is made infinitely large, that is the number of vertices N tends to infinity.

2 Generating Function Method

2.1 Undirected Graphs

In this section we outline the generating function method that is used to investigate these graphs. More extensive treatments can be found in [2] and [3] from where the formalism has been taken. Let $G_0(x)$ be the generating function for the probability districution that a vertex on the graph has degree k with probability p_k. This will be our most fundamental generating function.

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \tag{2.1}$$

$$G_0(1) = 1 \tag{2.2}$$

Here the second equation follows from the normalization of the probability distribution p_k. This property will continue to hold for all other derived generating functions unless otherwise specified.The kth probability can be determined by taking the kth derivatives of our generating function as follows.

$$p_k = \frac{1}{k!} \frac{d^k G_0}{dx^k}\bigg|_{x=0} \tag{2.3}$$

2.1.1 Moments

We can obtain information about the probability distribution by calculating the moments. The first moment for example gives us the average degree of a vertex in a graph having degree distribution p_k. Higher moments can be calculated by taking higher derivatives.

$$z = \langle k \rangle = \sum_k k p_k = G_0^{'}(1) \tag{2.4}$$

$$\langle k^n \rangle = \sum_k k^n p_k = \left[\left(x \frac{d}{dx} \right)^n G_0(x) \right]_{x=1} \tag{2.5}$$

2.1.2 The law of Powers

If the distribution of a property k of an object is given by a generating function, then the distribution of the total of k over m independent realizations of the object is generated by the mth power of the distribution. We can understand this by an illustrative example. Let us expand $G_0^2(x)$.

$$[G_0(x)]^2 = \sum_{jk} p_j p_k x^j x^k = p_0 p_0 + (p_0 p_1 + p_1 p_0)x + (p_0 p_2 + p_1 p_1 + p_2 p_0)x^2 + \cdots \tag{2.6}$$

Hence we see that the coefficient of the power of x^n correctly gives the probability that the sum of the powers of the vertices in n.

2.1.3 Other Properties

Let us determine the distribution of the degree of the vertices we arrive at by following a randomly chosen edge. Such an edge arrives at the vertex with probability proportional to kp_k. Hence the correctly normalized distribution is generated by

$$\frac{\sum_k k p_k x^k}{\sum_k k p_k} = x \frac{G_0^{'}(x)}{G_0^{'}(1)} \tag{2.7}$$

If we want to find the distribution for outgoing edges from the vertex we arrived at by following the randomly chosen edge, we divide by x to allow for the vertex we arrived by

$$G_1(x) = \frac{G_0^{'}(x)}{G_0^{'}(1)} = \frac{G_0^{'}(x)}{z} \tag{2.8}$$

The probability that each of these vertices connects with the original one goes as N^{-1}, which we neglect in the limit of large N, this limit as earlier commented is going to be very important in determining the position of the phase transition and its existence. Hence, making use of the power law, the distribution of the *number of second neighbours* is generated by

$$\sum_k p_k \left[G_1(x) \right]^k = G_0 \left(G_1(x) \right) \tag{2.9}$$

The average number of second neighbours are given by

$$z_2 = \left[\frac{d}{dx} G_0\left(G_1(x)\right)\right]_{x=1} = G_0^{'}(1)G_1^{'}(1) = G_0^{''}(1) \tag{2.10}$$

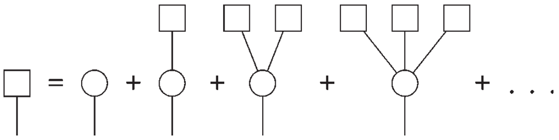

Figure 1: Schematic representation of the sum rule for connected components of vertices reached by following a randomly chosen edge

2.1.4 Component Sizes

Let $H_1(x)$ be the generating function for the distribution of component sizes which are reached by choosing some random edge and following it to its ends. We exclude from $H_1(x)$ the giant component which forms at phase transition, that is taken care of later on. Except at phase transition when a giant component appears a component has a finite size. The probability of a component having a closed loop goes as N^{-1} which we neglect in the limit of large N. The component can be imagined to be generated by the treelike structure as shown in Fig.(1), with either just the first vertex or the vertex with one or more component connected to it be edges each having the same distribution. Hence $H_1(x)$ must satisfy the consistency condition

$$H_1(x) \quad = \quad xG_1\left(H_1(x)\right) \tag{2.11}$$

Similarly, if we take a randomly chosen vertex the generating function for the component size at the end of each vertex is

$$H_0(x) = xG_0\left(H_1(x)\right) \tag{2.12}$$

In principle we can solve these equations given the probability distribution, but solving them analytically is almost impossible, they are generally, solved numerically.

2.1.5 Giant component and Phase transition

Although we cannot solve eqn's 2.11 and 2.12 analytically we can determine the mean component size

$$\begin{aligned}\langle s\rangle &= H_0^{'}(1) = 1 + G_0^{'}(1)H_1^{'}(1)\\ H_1^{'}(1) &= 1 + G_1^{'}(1)H_1^{'}(1)\\ \Rightarrow \langle s\rangle &= 1 + \frac{G_0^{'}(1)}{1 - G_1^{'}(1)} = 1 + \frac{z_1^2}{z_1 - z_2}\end{aligned}$$

We see that this expression diverges when $G_1^{'}(1) = 1$. This is the point of the Phase transition at which the giant component first appears.
When there is a giant component $H_0(1) < 1 = 1 - S$. Here, S is the fraction of the graph formed by the giant component. $S = 1 - G_0(u)$ where
$u = H_1(1)$ is the smallest non-negative real solution of $u = G_1(u)$. At this point the correct average size (barring size of the giant component) is

$$\langle s\rangle = \frac{H_0^{'}(1)}{H_0(1)} = 1 + \frac{zu^2}{(1 - S)(1 - G_1^{'}(u))} \tag{2.13}$$

Also, at the phase transition, Newman has shown in [3] by expanding H_1^{-1} around 1 that H_1 and H_0 both scale as $(1-x)^{\beta}$ for $x \to 1$ where $\beta = \frac{1}{2}$. He also shows that near the phase transition the probability distribution has the form $P_s \approx s^{-\alpha}e^{-\frac{s}{s*}}$ where s^* is a cutoff parameter related with the radius of convergence of the generating function. He also shows that $\alpha = 1 + \beta = \frac{3}{2}$. These results match with the results for the Poisson graph distribution.

2.2 Directed Graphs

The calculation for directed graphs is very similar so we just briefly look at the basic formulas. The generating function has now two indices j- the in-degree and k the outdegree.

$$G(x, y) = \sum_{j,k} p_{jk}x^j y^k \tag{2.14}$$

The indegree must be equal to the outdegree, hence,

$$\begin{aligned}\sum_{jk}(j-k)p_{jk} &= 0\\ \Rightarrow \left.\frac{\partial G}{\partial x}\right|_{x,y=1} &= \left.\frac{\partial G}{\partial y}\right|_{x,y=1} = z\end{aligned} \tag{2.15}$$

G_0 and G_1(F_0 and F_1) generate the number of outgoing(incoming)edges leaving(arriving) a randomly chosen vertex and the number leaving the vertex reached by following a randomly chosen edge. These functions are given as

$$\begin{aligned} F_0(x) &= G(x,1). \qquad F_1(x) = \frac{1}{z}\frac{\partial G}{\partial y}\bigg|_{y=1} \\ G_0(y) &= G(1,y). \qquad G_1(y) = \frac{1}{z}\frac{\partial G}{\partial x}\bigg|_{x=1} \end{aligned} \tag{2.16}$$

Other properties are derived as above. The internet with the pages as vertices and the links as edges is an example of a directed Graph. As before, the position of the phase transition is given by when $G_1'(1) = 1$. It is equally valid to say that the giant component occurs when $F_1'(1) = 1$.

3 Network structure of the Internet and some Results validating the above Formalism

There has been a lot of recent activity about the network structure of the internet. The results and the graphs referred to in this section are taken from [3] and [4] reproduced in [3]. Similar results are given in [7]. This data has been gathered by "crawls" of search engines across the internet. This means this data is accurate as far as the out degree is concerned however to get an accurate estimate of the in degree we would have to crawl the whole WWW, which is impractical. However certain results which depend only on the out degree are tested. First let us show that the data collected is evidence of the fact that real life data about random graphs is not Poisonnian but in the case of WWW it has a Power Law behaviour. The distributions are well plotted using a power law of form $p_k = C(k + k_0)^{-\tau}$. [4] explains this using a combination of stochastic dynamical growth and a difference in ages and growth rates of the sites. 3 shows clearly the disparity in the in and out degrees due to the mode of data collection. Using the generating function formalism described in the previous section, we determine the fraction of the graph S which is reached by the giant strongly connected component. This is given as $S_{in} = 1 - G_0(1 - S_{in})$. Thus $(1 - S_{in})$ is a fixed point of $G_0(x)$. Using numerical simulation we find that S_{in} is 53%. This matches closely with the direct measurements reported by Broder et. al. which says that 49% of the web falls in the giant component. The simulations can be done in two ways. In the first way we can generate a particular degree sequence for the graph by taking N integers at random from the probability distribution by which the vertex degrees are governed. Then we average over the results

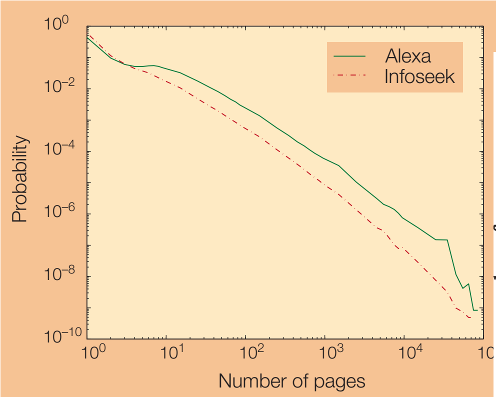

Figure 2: Log-Log plot of Data from Alexa and InfoSeek Crawls which cover 259,794 and 525,882 pages respectively. The τ value reportedwith 95% confidence is [1.647,1.853] and [1.775,1.909]

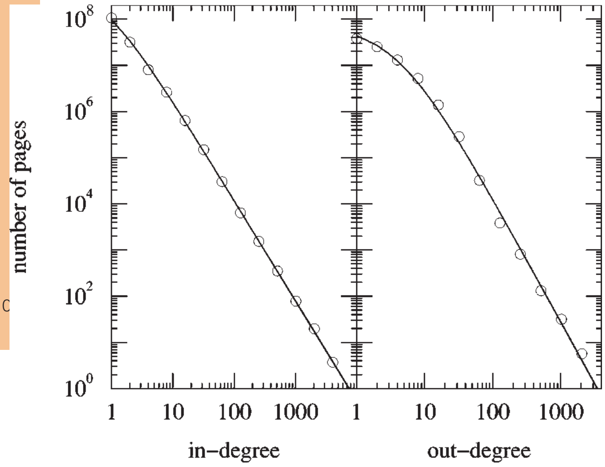

Figure 3: Probability distribution of in and out degree from [3]τ values are 2.17 and 2.69 from in and out degree distributions

of all graphs that can be obtained using that degree sequnce. This we average over a large number of implementations of the degree sequence. So, this is akin to the canonical ensemble in Statistical mechanics. In contrast we, can choose a particular probability distribution for degrees and choose N numbers from that distribution provided that N is very large. Then we just average over all graphs generated by that degree distribution. This is somewhat like the microcanonical ensemble.**Although it seems to be a reasonable validation of the formalism, one must note that in the simulations performed by [3] the assumption was made that** $p_{jk} = p_j p_k$, i.e. the in and out degrees are correlated, something there is no reason to believe.

4 Modification of Model to include Percolation

In the introduction we have already motivated the need to introduce percolation to study network resilience in random graphs. The introduction of percolation just modifies certain terms in the generating function formalism. First let us briefly go over them. Here we concentrate only on undirected graphs. The discussion and results in this section follow closely[1]. We begin by examining site percolation when occupation probability is an arbitrary function of vertex degree. So, p_k is the probability that a randomly chosen vertex has degree k and q_k is the probability that that it is occupied. Then the generating function is given by

$$F_0(x) = \sum_{k=0}^{\infty} p_k q_k x^k \tag{4.1}$$

As before the distribution for outgoing edges reached from a given vertex has a probability distribution generated by

$$F_1(x) = \frac{\sum_k k p_k q_k x^{k-1}}{\sum_k k p_k} = \frac{F_0^{'}(x)}{z} \tag{4.2}$$

Now, $H_1(x)$ is the generating function that one end of a randomly chosen edge leads to a percolation cluster of a given number of occupied vertices. The cluster has zero vertices if the vertex at the end of the edge in question is unoccupied. This happens with probability $1 - F_1(1)$, or it can have k other edges leading out of it with a distribution $F_1(x)$. Thus $H_1(x)$ satisfies a consistency relationship given below. The probability distribution for the size of a cluster to which a randomly chosen vertex belongs is generated by

$$\begin{aligned} H_1(x) &= 1 - F_1(1) + xF_1\left[H_1(x)\right] \\ H_0(x) &= 1 - F_0(1) + xF_0\left[H_1(x)\right] \end{aligned} \tag{4.3}$$

For the case of joint site/bond percolation with uniform site and bond percolation probability q_s and q_b the generating functions are

$$\begin{aligned} H_1(x) &= 1 - q_s q_b + q_s q_b x G_1\left[H_1(x)\right] \\ H_0(x) &= 1 - q_s + q_s q_b x G_0\left[H_1(x)\right] \end{aligned} \tag{4.4}$$

where $G_0(x) = \sum_k p_k x^k$ and $G_1(x) = G_0^{'}(x)/z$ are the generating functions for vertex degree alone as before. Note that we get the results of the previous sections by setting q_s and q_b to 1. Let us try to analyze uniform site occupation probability with q, i.e. we set $q_s = q$ and $q_b = 1$. Again as in previous sections, we get for mean cluster size

$$\begin{aligned} \langle s \rangle = H_0^{'}(1) &= q + qG_0^{'}(1)H_1^{'}(1) \\ &= q\left[1 + \frac{qG_0^{'}(1)}{1 - qG_1^{'}(1)}\right] \\ q_c &= \frac{1}{G_1^{'}(1)} \end{aligned} \tag{4.5}$$

Here q_c marks the point at which the percolation phase transition happens and the giant component first forms. In the language of disease propagation this is the point at which the epidemic first happens. More extensive work in this direction has been done by Newman in [8]. For networks this marks where the network achieves large scale connectivity and can work as an effective distribution network. As shown before the Internet has a power law distribution for vertices. It seems several other networks like the collaborations of scientists also have a similar distribution but with an exponential cutoff as given below.

$$\begin{aligned} p_k &= 0 \text{ for } k = 0 \\ &= Ck^{-\tau}e^{-\frac{k}{\kappa}}\text{for } k \geq 1 \end{aligned}$$

The Internet can then be taken as a particular case with $\kappa \to \infty$.This data is now used for simulations(The exponential cutoff is included so that the generating function is finite.). The results of these calculations are reproduced below in Figure. 4. We see that for $\tau = 2.5$(close to the exponent for Internet Data) and $\kappa = 100$ $q_c = 0.17$. This means more than 80% of the nodes need to be removed before we can destroy large scale connectivity.

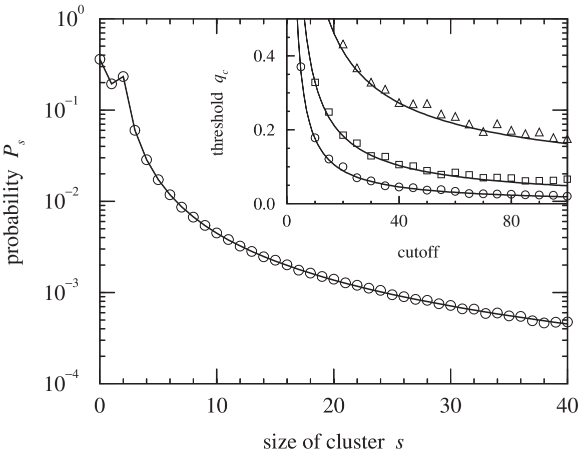

Figure 4: Probabilty that a randomly chosen vertes belongs to a cluster of s sites for $\kappa = 10, \tau = 2.5, q = 0.65$. Percolation threshold for $\tau =$ 1.5(circles),2(squares),2.5(Triangles)

Another question with recent interest is the connectivity of the network when nodes having high degree are progressively removed. Let us suppose we remove all nodes having degrees greater than k. This has been studied in [5] and [6].

$$q_k = \theta(k_{\max} - k) \tag{4.6}$$

The size of the giant component S is determined using the following equations

$$\begin{aligned} S = 1 - H_0(1) &= F_0(1) - F_0(u) \\ u &= 1 - F_1(1) + F_1(u) \end{aligned} \tag{4.7}$$

. The results of the simulation for pure power law distributions are as shown in Figure.5 From the top part of 5 we see that only a small portion of the vertices (less than 3%) needs to be removed to destroy the giant component. But from the bottom part we realize that k_{max} must be really small, so that all vertices with degree greater than almost 10 must be removed for $\tau = 2.7$. In this sense the network is stable.

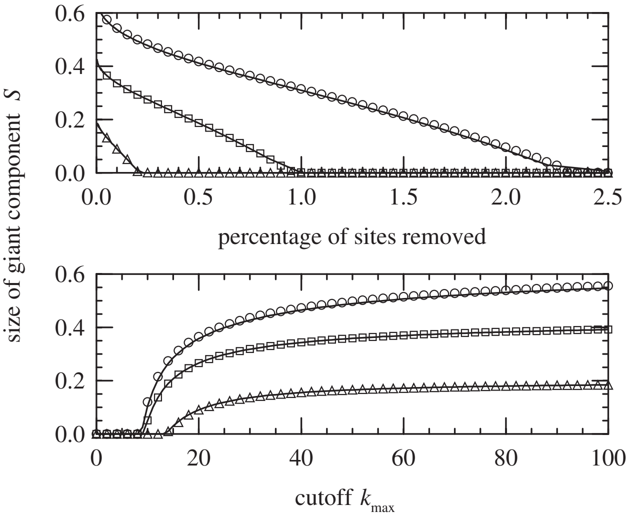

Figure 5: Size of the giant component with power law distribution τ =2.4(circles), 2.7(squares)3(triangles)

5 Conclusion

In the report we have followed some work done regarding the network structure of graphs which are completely random in every aspect barring their degree distributions. We look over both directed and undirected graphs and try to obtain results using generating function methodology. The results from simulations of the equations agree with analytical calculations. However, there is not much experimental data to validate the above models. The

only available data is regarding the the size of the giant component in WWW, which is reported in Section 3. However, there are two problems regarding this data. One, we do not have enough information about the in -degrees. Two, we assume that the in and out degrees are uncorrelated. Also, from the point of view of disease propagation, we should note the that study is primarily kinetic, in the sense that we determine the structure of the network at the point of spread of an epidemic. However, for the study to be useful it must be dynamic too in the sense that we must determine the evolution of the structure of the network in terms of experimentally determinable quantities. Such models exist in the way of SIR and SIS models, and such studies have been done in [8] and are reviewed in [2].
Such problems are very important because thay provide us with tools with which to analyze networks which are of everyday use to us, like telecommunication or airline routes. These tools will enable us to design robust networks which would still have connectivity with respect to removal of nodes, accidental or for example by sabotage. However such studies would be helped by collection of new data.

References

[1] Callaway, Newman, Strogatz, Watts.(Phys. Rev. Lett) *Network Robustness and Fragility: Percolation on Random Graphs* **85**, 5626(2000.)

[2] M.E.J. Newman(SIAM) *The structure and function of complex networks* **45** 167 (2003) .

[3] Newman, Strogatz, Watts.(Phys. Rev. E.) *Random graphs with arbitrary degree distributions and their applications.* **64** 026118(2001).

[4] Huberman, Adamic(Nature)*Internet: Growth dynamics of the World-Wide Web* 401, 131(1999).

[5] Albert,Jeong,Barabasi(Nature)*Error and attack tolerance of complex networks* **406**, 378(2000).

[6] Broder, Kumar, Maghoul,et.al.(Computers Networks)*Graph Structure in the Web* **33**, 309(2000).

[7] M.Faloutsos,P.Faloutsos,C.Faloutsos(Comp. Comm. Rev)*On power-law relationships of the Internet topology* **29**, 251 (1999).

[8] M.E.J. Newman(Phys. Rev. E)*Spread of Epidemic Disease in Network* **66**,06128(2002).