Global Properties of Some Preferential Attachment Graphs

Julie Lang

May 9, 2011

Abstract

We are looking at a model of preferential attachment graphs that can evolve in five different ways. On any given iteration, a single vertex may be connected to the graph, a path may be connected to the graph, an edge may be added to the graph, an edge may be removed from a graph, or a vertex may be removed from a graph, with each option having a given probability of occurring. If a single vertex or a path is added, one or both of the new vertices also may be connected to the neighborhood of the existing vertex. We explore the order, size, and degree distribution of the graphs which result from a large number of iterations.

Acknowledgment Special thanks to Angela Collier and Blake Nickell for all of their help with this project, and to Dr. R. Duane Skaggs for being an excellent advisor.

1 Introduction

How can the World Wide Web be related at all to the neurons in your brain? What similarities could possibly exist between Hollywood actors and atoms in a Bose-Einstein condensate? At first glance, many interesting systems like these may seem to have very little in common with each other. However, when inspected in a specific way, all of these show the same basic structure. They are complex networks and can be described using graph theory.

We begin with a few definitions. A *graph* is defined as a finite nonempty set V together with an irreflexive, symmetric relation R on V. The set of

symmetric pairs in R is denoted by E. Each element of the set V is called a *vertex* and each element of E is called an *edge* [5]. Graphs can be represented pictorially where the vertices are shown as dots and edges are shown as lines that connect those dots.

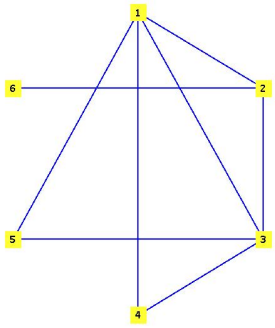

Figure 1: An example of a graph.

The *order* of a graph is $|V|$, describing the number of vertices in the graph. Similarly, the *size* of a graph is $|E|$, describing the number of edges. If an edge $(u, v) \in E$, u and v are said to be *adjacent* vertices and the edge (u, v) is said to be *incident to* u and v. The number of edges incident to a vertex v is called the *degree of* v [5]. We will denote the degree of a vertex v by k_v. See Chartrand [5] for any additional definitions that are not included in this paper.

When graphs are used to represent complex networks, the vertices represent elements that comprise the system and the edges represent the interactions between those elements. For example, a vertex may represent an airport and an edge may represent a flight from one airport to another. Using graphs to analyze complex networks allows results to relate to many different disciplines, from biology to sociology to entertainment.

It has been observed that complex networks tend to follow a power law. That is, the probability that any given vertex in a system has degree k follows $P(k) \sim k^{-\beta}$, where β is a positive constant for that particular system. Barabási and Albert [1] give a few examples of systems and their respective β values. In the actor collaboration graph, a vertex represents an actor. Two vertices are connected by an edge if those actors have been in the same movie. They report that $\beta_{actor} = 2.3$. Citations in science also form a complex network where a vertex is a published paper and there is an edge

between two vertices if one paper cites the other. It turns out that $\beta_{cite} = 3$. This power law is related to the 80/20 rule, where approximately 80% of goods are owned by 20% of a population [2]. This distribution allows for the existence of a few highly connected vertices, but the graph is predominately comprised of vertices that are connected by only a few edges.

Also, the networks in question are quite large; they may have thousands, millions, or even billions of vertices. But the power law distribution of the degrees remains, regardless of how big the system is. For this reason, it is frequently called scale-free. The scale-free power law is a distinguishing feature of complex networks. It is important that a model is able to produce this distribution.

Another feature of complex networks is known as the clustering effect. That is, two vertices are more likely to be connected if they are both adjacent to the same vertex [6]. For example in social networks, a person is more likely to befriend someone if they share a common friend. In the World Wide Web, two websites that linked to the same document are more likely to link to each other, too. This clustering is very clear in some networks. For example, the actor collaboration network demonstrates this clustering very well since many actors are associated with a particular movie genre (such as comedy or horror). So an actor will be more likely to be in the same move as another actor if they are associated with the same genre [2].

In this paper, we present many models that can describe these complex networks. We describe how subsequent models change to build upon the ideas of previous models. Following suit, we present a new model with additional features. We will focus on the global properties of these models, such as order, size, average degree or the distribution of degrees. But let us begin with begin with the classical model.

2 The Erdős-Rényi Model

Initially, complex networks were described by the Erdős-Rényi random graph model. This model creates a graph $G_{n,p}$ based on two parameters: a fixed number of vertices n and a probability p. Each possible edge has the probability p of being in the graph. The resulting graph $G_{n,p}$ will have different properties with different values of p. Chung and Lu [6] summarize the six ranges of p that produce these different features in the graph:

- When $p = o(1/n)$, the graph $G_{n,p}$ is a disjoint union of trees, where

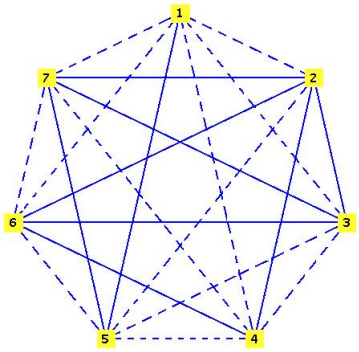

Figure 2: In the Erdős-Rényi model, all of the edges had equal probability of appearing, but only some of them actually end up in the final graph.

a *tree* is defined as a graph that has no cycles [5]. In this range of p, trees on c vertices (for an integer $c \geq 3$) will only appear when p is on the order of $n^{-c/(c-1)}$.

- When $p \sim c/n$ where $0 < c < 1$, the graph $G_{n,p}$ consists of either trees or unicyclic components (trees with an additional edge that creates a single cycle). Formally, a *component* is a connected subgraph, where a subgraph is *connected* if a path exists between any two vertices [5]. Almost all vertices are in components which are trees. The largest connected component is a tree with about $(c - 1 - \log c)^{-1}(\log n - \frac{5}{2} \log \log n)$ vertices.

- When $p \sim 1/n + \mu/n$ is the range in which the "giant" component emerges. This range is sometimes called "the double jump". For $\mu < 0$, all components are trees or unicyclic and most vertices are in components of size $O(1)$ and the largest component has size $O(\log n)$. When $\mu = 0$, the largest component has size on the order of $n^{2/3}$. When μ is positive, there are still a number of small components of size $O(\log n)$, but most of the vertices are contained in a unique giant component, which has size $O(n)$.

- For $p \sim c/n$ where $c > 1$, there is a single giant component and all other components are mostly relatively small trees. About $n - f(c)n + o(n)$ of the vertices are in these small trees and approximately $f(c)n$ vertices

are in the giant component, where

$$f(c) = 1 - \frac{1}{c}\sum_{i=1}^{\infty} \frac{i^{i-1}}{i!}(ce^{-c})^i.$$

- For $p = c \log n/n$ with $c \geq 1$, the graph $G_{n,p}$ is almost surely connected.
- For $p \sim \omega(n) \log n/n$, the graph $G_{n,p}$ is almost surely connected and the degrees of almost all vertices are asymptotically equal. See [6] for a definition of $\omega(n)$.

However, the degrees of the vertices in a graph produced in the Erdős-Rényi model follow a Poisson distribution. That is, as stated in [1], the probability that a vertex has degree k is $P(k) = \frac{e^{-\lambda}\lambda^k}{k!}$ where

$$\lambda = n\binom{n-1}{k}p^k(1-p)^{n-k-1}.$$

A scale-free power law is quite different, so the Erdős-Rényi model does not capture some important properties of these complex networks. It turns out that the scale-free power law can be brought about by employing two mechanisms: growth and preferential attachment.

3 Growth and Preferential Attachment

Simply put, preferential attachment is the idea that a vertex is more likely to connect to highly connected vertices. Just as the websites with many links tend to get more links and the people with a lot of money are the ones that get even more money, the vertices with many incident edges are the ones that tend to get more edges. Preferential attachment is a rich-get-richer phenomenon. It is a local mechanism that changes the global properties of the graph [1, 2].

Preferential attachment is important in constructing graphs that model complex networks, but so is growth. Barabási and Albert [1] determined that one or the other is not sufficient to produce the scale-free power law; both mechanisms should be present. In one model, they include growth without preferential attachment. At each time step t, a new vertex is introduced into the graph and connected to an existing vertex chosen at random. This is

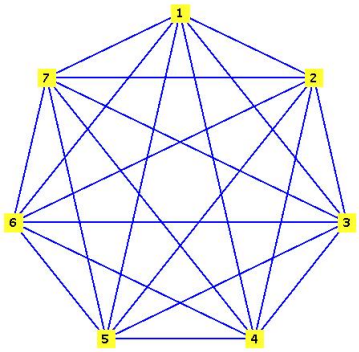

Figure 3: The complete graph on seven vertices, K_7.

reported to result in a degree distribution that follows a decaying exponential, $P(k) \sim e^{-\alpha k}$, rather than a power law.

A second model incorporates preferential attachment without growth. Beginning with n vertices, each time step adds an edge between two chosen vertices with

$$P(k_i) = \frac{k_i}{\sum_j k_j}$$

being the probability of choosing a vertex i. The degree distribution of this model initially resembles a power law, but this changes as t grows. Since n is unchanged, at $t \simeq n^2$, every two of the vertices are adjacent. This is called a *complete graph* and is denoted K_n [5]. An example of a complete graph is shown in the figure. After every edge is included, the graph remains static.

Since both of these models fail to produce the scale-free power law, Barabási and Albert [1] concluded that both growth and preferential attachment are essential for these graphs to describe the real-world systems with which we are concerned.

Theorem 1. *A scale-free power law distribution can result from implementation of both growth and preferential attachment.*

Since the scale-free power law is desired in modeling complex networks, growth and preferential attachment are extensively employed in existing models. There are many different ways to incorporate these mechanisms into a model and their implementation affects the global properties of the resulting graph.

4 Existing Dynamic Models

4.1 The Barabási-Albert model

Unlike the Erdős-Rényi model, the Barabási-Albert model [1] is dynamic. It begins with a small number (n_0) of vertices and a small number (m_0) of edges. At each time step t, a new vertex is introduced to the system and connected to r other vertices. These vertices are chosen randomly with the probability P of choosing a vertex i depending upon the degree of i. That is,

$$P(k_i) = \frac{k_i}{\sum_j k_j}.$$

Notice that this simple rule incorporates both growth and preferential attachment into the model. The resulting degree distribution indeed follows a scale-free power law. When $r = 5$, β is observed to be 2.9. This is very similar to the exponent of some of the real-life networks that are discussed in [1]. Since each time step adds exactly one new vertex to the graph, the expected order at time t is $N(t) = n_0 + t$. Each vertex is then connected to exactly r vertices. This introduces r edges at each time t, so the expected size of the graph at time t is $M(t) = m_0 + rt$.

4.2 The Bollobás-Riordan Model

Bollobás and Riordan proposed a very similar model that eliminates the need for an n_0 and an m_0 [3, 4]. The initial graph contains no vertices. At each time step t, a new vertex is added. Then r edges are introduced, all of which have the new vertex as one endpoint. The other endpoint will be either an existing vertex or the new vertex itself. In this way, the Bollobás-Riordan model allows for loops and multiple edges. For $r = 1$, the probability that the new vertex will connect to a vertex i is

$$P(k_i) = \begin{cases} \dfrac{k_i}{2t-1} & 1 \leq i \leq t-1 \\ \dfrac{1}{2t-1} & i = t \end{cases}.$$

In this model, a loop increases the degree of a vertex by 2. So the $2t-1$ in the denominator is actually the sum of the degrees of all of the vertices at time $t-1$, plus the end of the new edge that is certain to be incident to the new

vertex. Observe that as t becomes large, the probability that the new vertex connects to itself becomes quite small. Therefore, loops become uncommon and thus do not contribute much to the global properties of the graph. The only difference between this model and the Barabási-Albert model is which vertices the new vertex can connect to, so the expected order is $N(t) = t$ and the expected size is $M(t) = rt$. Since the initial graph contains no vertices, n_0 and m_0 are both zero.

4.3 Chung-Lu Models

Chung and Lu [6] present many models for complex networks. Only a few are described below, those that are most closely related to our project.

A growth-deletion model is presented in which at each time t, one of four things will happen.

- With probability p_a, a new vertex is added to the graph and connected to an existing vertex that is selected with a probability proportional to its degree.
- With probability p_b, a random edge is added between two existing vertices chosen with a probability proportional to the degree of each.
- With probability p_c, a vertex is chosen at random and is deleted, along with all of its incident edges.
- With probability p_d, a randomly chosen edge is deleted.

The assumption here is that $p_a > p_c$ and $p_b > p_d$ so that the graph grows. This model is a little more cognizant of how some real networks evolve. Many real-world systems allow their elements to be removed from the system or to interact with other preexisting elements. This model almost surely displays the scale-free power law, with $\beta = 2 + (p_a + p_b)/(p_a + 2p_b - p_c - 2p_d)$. Recall that this remains constant throughout the evolution of the graph. Also, the expected average degree is given as $(p_a + p_b - p_d)/(p_a + p_c)$.

Another interesting model proposed by Chung and Lu [6] is the duplication-deletion model. This begins with some initial graph G_0. For each time step t, one of two things will occur. With probability r, a vertex is "duplicated". That is, a vertex is selected with probability p and is connected to a new vertex. This new vertex is then connected to the neighborhood of the selected vertex. This means the new vertex is connected to each neighbor of

the selected vertex. Hence the selected vertex is duplicated, since it is now indistinguishable from the new vertex. However, for times when a duplication step is not carried out, (meaning that for each time step t, with probability $1 - r$) a random vertex is selected and deleted, along with all of its incident edges. It is assumed that $r > 1/2$ so that the graph grows with time. This model results in a power-law degree distribution where β satisfies

$$(\beta - 1)(1 + p - \frac{1}{r}) = 1 - p^{\beta - 1}.$$

The duplication-deletion model focuses heavily on the clustering effect that is observed in many of the networks that we are trying to describe, since it adds edges to the neighborhood of a vertex.

Even more closely related to the clustering effect is Chung and Lu's mixed duplication model [6] that does not allow for deletion. This model begins with some initial graph G_0 and at each time step t, one of two things will occur. With probability q, a new vertex is connected to an existing vertex that is selected randomly and uniformly from G_{t-1}. Then for each neighbor of the selected vertex, with probability p, the neighboring vertex is connected to the new vertex. For all other times (that is, for each time step t, with probability $1 - q$), a new vertex is connected to an existing vertex that is selected randomly and uniformly from G_{t-1}. Then edges are added between the new vertex and every neighbor of the selected vertex. Note that this model does not incorporate preferential attachment; the vertices all have an equal chance of being selected. However, this still generates a power law distribution with an exponent β that satisfies

$$\beta(1 - q) + pq(\beta - 1) = 1 - qp^{\beta - 1}.$$

5 Off-line and On-line Models

Chung and Lu [6] distinguish between two distinct types of models. An off-line model has a fixed number of vertices. A graph is grown only by the addition or deletion of edges while the number of vertices is held constant. The Erdős-Rényi model is an example of an off-line model. On-line models are dynamic and so the order and size of the graph can change at each time step that is taken, according to some rule. Notice that growth is explicitly incorporated into on-line models. One may think of an on-line model as a

sequence of off-line models where the graph at time t may depend on the graph at a previous time. Clearly, an on-line model is more difficult to analyze than an off-line model, but on-line models come closer to reflecting the evolution of the complex networks that we see in the real world.

Nonetheless, all of the aforementioned models focus heavily on the end result. They look at the properties that a graph will have after some large number of iterations and do not place too much emphasis on how realistic the growth of that graph was. In the spirit of modeling the processes used in the evolution of these systems, we propose an on-line random graph model that grows with a few additional features.

6 Description of Our Model

Our model is slightly more involved than the other models that have been presented. We are interested in modeling a system like Facebook. In this case, a vertex is a Facebook profile and two vertices are connected by an edge if the corresponding profiles are friends on Facebook. We assume that when Facebook was first created, there was a very small number of profiles and each profile was friends with every other profile. Therefore, our model begins with a complete graph K_4. At each time step t, one of five things will happen.

- With probability p_1, a single vertex will be added and connected to an existing vertex chosen at random.
- With probability p_2, two vertices joined by an edge, called a path P_2, will be added and one of the vertices (called the *strong vertex*) will be connected to an existing vertex chosen at random.
- With probability p_3, an edge will be added between two randomly chosen nonadjacent vertices.
- With probability p_4, a randomly chosen edge will be deleted.
- With probability p_5, a randomly chosen vertex will be deleted along with all incident edges.

Our motivation for adding a P_2 was to model a parent and a child joining Facebook. The strong vertex represents the child's profile and the parent's

profile is represented by the *weak vertex*, the remaining vertex in the P_2. Note that whenever a P_2 is introduced, the strong vertex always connects to a selected vertex in the existing graph. This is consistent with the child becoming Facebook friends with someone who already has a Facebook profile.

Additionally, a common rule in households is that a child may use Facebook only if that child adds their parent and the parent can be friends with all of the child's friends. That idea involves two connected vertices being introduced at the same time (which is why we add a P_2). But in order to model the parent adding all of the child's friends, we must have some additional parameters.

- If a P_2 is added, let h_0 be the probability that the weak vertex also connects to the selected vertex.
- If a P_2 or a single vertex is added, let h_1 be the probability that the (strong) vertex connects to the neighborhood of the selected vertex.
- If a P_2 is added, the weak vertex has connected to the selected vertex, and the strong vertex has connected to the neighborhood of the selected vertex, let h_2 be the probability that the weak vertex connects to the neighborhood of the selected vertex.

For clarification, the reason h_1 applies to a single vertex or a P_2 is because the strong vertex should act independently of the weak vertex. Essentially, the strong vertex should behave as if it were a single vertex. Therefore, both should have the probability h_1 of connecting to the neighborhood of the selected vertex. Also, it only makes sense for the weak vertex to connect to the neighborhood if it has connected to the selected vertex itself and the strong vertex has connected to the neighborhood of the selected vertex. Both conditions must be met in order for the weak vertex to connect to the neighborhood. In this sense, the behavior of the weak vertex is entirely dependent upon the behavior of the strong vertex.

7 Expected Order and Size of Our Model

Naturally, the equations for the expected order and size for our model will be a bit involved since ours is more complicated than the other models that have been presented. Nevertheless, we have developed formulas that give the expected order and size of our model at any given time t.

7.1 Expected Order

Proposition 1. *The expected order of our model at time t is given by $N(t) = 4 + t(p_1 + 2p_2 - p_5)$.*

Proof. Initially, we have a K_4, so $N(0) = 4$. After that, p_1, p_2, and p_5 are the only probabilities associated with the number of vertices. The P_2 step adds two vertices instead of one and the deletion step removes one vertex. We multiply the number of vertices added in a given step by the probability of that step occurring to get $p_1 + 2p_2 - p_5$. Since this quantity remains unchanged for each time step t, the expected order is

$$N(t) = 4 + t(p_1 + 2p_2 - p_5).$$

□

7.2 Expected Size

Proposition 2. *The expected size of our model at time t is given by*

$$M(t) = M(t-1)+p_1+(2+h_0)p_2+p_3-p_4+[h_1(p_1+p_2+p_2h_0h_2)-p_5]\frac{2M(t-1)}{N(t-1)}.$$

Proof. Before taking a time step t, the number of edges in the graph is denoted $M(t-1)$. Therefore to find the size after t time steps, $M(t)$, we need only add the value $M(t-1)$ to the expected number of edges introduced into the graph from an individual time step. There are two portions of the time step that will added edges.

Portion 1: Addition of vertices or explicit addition or removal of edges

We expect one edge to be added with the probability p_1. This edge connects a single vertex to a selected vertex. With probability p_2, two edges will be added. One of these connects the strong vertex to the weak vertex and the other connects the strong vertex to a selected vertex. Additionally, with probability p_2h_0, the weak vertex connects to the selected vertex. One edge will be added with probability p_3 from connecting two existing nonadjacent vertices and -1 edges will be added with probability p_4 when a random edge is deleted. Summing all of these contributions, we find that the expected number of edges to be contributed from the first portion of a time step is $p_1 + (2 + h_0)p_2 + p_3 - p_4$.

Portion 2: Connecting a new vertex to a neighborhood or deleting a vertex

The average degree of a vertex must be considered when a vertex connects to a neighborhood and when a vertex is deleted (since all of the incident edges are deleted along with the vertex). Because loops and multiple edges do not occur in our model, each incident edge connects a vertex to a unique neighbor. The average degree of a vertex at a time t is given by $2M(t)/N(t)$. Therefore, when a vertex connects to a neighborhood or when all the incident edges of a vertex are deleted, $2M(t-1)/N(t-1)$ vertices are added or removed respectively. Now we must consider the probability of connecting a vertex to a neighborhood. A single vertex may connected to the neighborhood of the selected vertex with probability p_1h_1. The probability that a P_2 is added and the strong vertex is connected to the neighborhood of the selected vertex is p_2h_1. Due to its dependence on the strong vertex, the probability that the weak vertex is connected to the neighborhood of the selected vertex is given by $p_2h_0h_1h_2$. Finally, all of the incident vertices of an existing vertex are removed with p_5. Summing these probabilities and multiplying by the expected average degree of a vertex in the existing graph (that is, at time $t-1$) gives $[h_1(p_1+p_2+p_2h_0h_2)-p_5](2M(t-1))/N(t-1)$.

Adding the number of edges in the existing graph and those introduced by portions 1 and 2 of an individual time step, we find

$$M(t) = M(t-1)+p_1+(2+h_0)p_2+p_3-p_4+[h_1(p_1+p_2+p_2h_0h_2)-p_5]\frac{2M(t-1)}{N(t-1)}.$$

□

Due to the complexity of this model, two formulas for the expected size were developed. It can occasionally be more helpful to use the equivalent expression below.

Proposition 3. *The expected size of our model at time t is given by*

$$\tilde{M}(t) = 6+t[p_1+(2+h_0)p_2+p_3-p_4]+[h_1(p_1+p_2+p_2h_0h_2)-p_5]\sum_{i=0}^{t-1}\frac{2\tilde{M}(i)}{N(i)}.$$

Proof. To show that $\tilde{M}(t)$ gives the expected size at a time t, we need only prove it is equivalent to $M(t)$. We proceed by induction.

Base case: $t = 1$

By definition of $\tilde{M}(t)$,

$$\tilde{M}(1) = 6+(1)[p_1+(2+h_0)p_2+p_3-p_4]+[h_1(p_1+p_2+p_2h_0h_2)-p_5]\sum_{i=0}^{(1)-1}\frac{2\tilde{M}(i)}{N(i)}.$$

$$\tilde{M}(1) = 6 + [p_1 + (2 + h_0)p_2 + p_3 - p_4] + [h_1(p_1 + p_2 + p_2h_0h_2) - p_5]\frac{2\tilde{M}(0)}{N(0)}.$$

The value of both $M(0)$ and $\tilde{M}(0)$ is defined to be 6, since our model begins with a K_4. Substituting in, we find

$$\tilde{M}(1) = M(0) + [p_1 + (2 + h_0)p_2 + p_3 - p_4] + [h_1(p_1 + p_2 + p_2h_0h_2) - p_5]\frac{2M(0)}{N(0)}.$$

Since $0 = t - 1$, the right-hand side is exactly the value of $M(1)$. So this implies $\tilde{M}(1) = M(1)$.

Induction step:

$$\tilde{M}(t+1) = 6 + (t+1)[p_1 + (2 + h_0)p_2 + p_3 - p_4]$$

$$+[h_1(p_1 + p_2 + p_2h_0h_2) - p_5]\sum_{i=0}^{t}\frac{2\tilde{M}(i)}{N(i)}$$

Distributing through the second term and pulling the last term out of the sum, this implies

$$\tilde{M}(t+1) = 6 + t[p_1 + (2 + h_0)p_2 + p_3 - p_4] + [h_1(p_1 + p_2 + p_2h_0h_2) - p_5]\sum_{i=0}^{t-1}\frac{2\tilde{M}(i)}{N(i)}$$

$$+[p_1 + (2 + h_0)p_2 + p_3 - p_4] + [h_1(p_1 + p_2 + p_2h_0h_2) - p_5]\frac{2\tilde{M}(t)}{N(t)}.$$

Notice that the first three terms of this expression are exactly the value of $\tilde{M}(t)$. Since the induction hypothesis states that $\tilde{M}(t) = M(t)$, this implies

$$\tilde{M}(t+1) = M(t) + p_1 + (2 + h_0)p_2 + p_3 - p_4 + [h_1(p_1 + p_2 + p_2h_0h_2) - p_5]\frac{2M(t)}{N(t)}.$$

$\tilde{M}(t+1) = M(t+1)$ □

7.3 Particular Results

Although the graphs produced by our model are random, we can use these formulas to anticipate what the resulting graph will be for certain cases. This has been helpful in checking the program that is mentioned in the next section.

There are many ways for our model to produce a complete graph. In a complete graph, the union of any selected vertex with its neighborhood will be the vertex set of the graph. Therefore, whenever a new vertex connects to the neighborhood of a vertex selected from a complete graph, it will connect to every vertex in the existing graph. This will result in another complete graph. Thus we can ensure that a graph produced from our model is complete at any time t. We need only ensure that each time a new vertex is introduced, it always connects to the neighborhood of the selected vertex and we cannot allow single edges to be deleted. That is to say that h_0, h_1 and h_2 must all equal 1 and $p_4 = 0$. Note that we do not prohibit the deletion of a random vertex. This is because when a vertex is deleted, only the edges that are incident to it will be removed. Since all of the edges between the remaining vertices are left untouched, the graph is still complete after a vertex is removed.

By inspecting the formulas for the expected order and size, one must note that a negative value implies that we expect an empty graph. Although one may expect an empty graph any time $N(t)$ is zero, it takes more severe restrictions to guarantee that the graph is empty. Clearly, to get an empty graph, t must be greater than 4, since it takes at least that long to remove all four vertices from the initial graph. Additionally, p_5 must be the only non-zero probability (discounting h_0, h_1, and h_2, since these only matter when p_1 or p_2 are non-zero). In this case, the graph will be empty for all $t \geq 4$.

8 Simulations

We are able to compare these expected values with experimental results. A program has been written in Maple that will grow these random graphs according to our model. It outputs many of the global properties that we are concerned with. We can analyze the output of this program and show specific examples of graphs constructed according the rules in our model.

For these first simulations, we will use a certain set of probabilities that seem reasonable for a system such as Facebook. These are $p_1 = 0.4, p_2 = 0.35, p_3 = 0.15, p_4 = 0.05, p_5 = 0.05, h_0 = 0.5, h_1 = 0.25, h_2 = 0.2$. A specific example of a graph grown according to these probabilities is shown in the figure.

The expected order at $t = 30$ is 41.5 and the expected size is 60.0. Running the program 20 times resulted in a mean order of 35.1 with standard

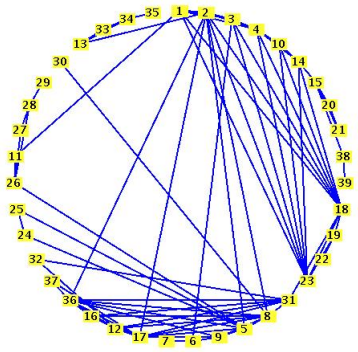

Figure 4: At $t = 30$, this graph has order 39 and size 81.

deviation 5.5 and a mean size of 61.9 with standard deviation 18.9. Due to the random nature of these graphs, the order and size will vary quite a bit from run to run, particularly when t is lower. For example, in these 20 runs, the size ranged from 27 to 111! So one should not be surprised that the standard deviation is rather large compared to the mean. A plot of the degrees for the example above is also provided.

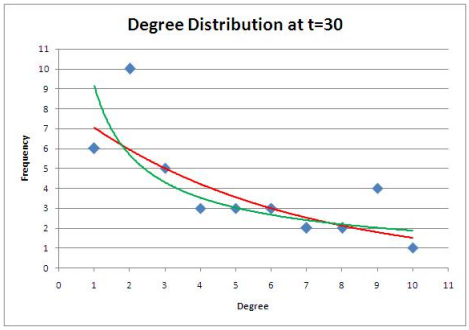

Figure 5: The degree distribution of a graph at $t = 30$ with a fitted power law (shown in green) and exponential (shown in red)

Fitting curves to the data, we see that neither the power law nor the exponential are a particularly good approximation. To evaluate the goodness of fit, we look at the value of R^2. A value of R^2 that is closer to 1 indicates a better fit. The values of R^2 for these are only 0.608 and 0.649 respectively.

The difficulty in fitting curves these curves is likely due to the small value of t.

Due to time constraints, only two runs were done at $t = 300$. The order for the first was 318 and its size was 603. The second had order 334 and size 579. The expected order and size were 379.0 and 545.0 respectively. One must keep in mind that these are only two specific examples, so differences between the expected values and the actual values are anticipated. The degree distribution is of particular interest in this case. The degree sequences of these examples are quite interesting. Plots of the degrees of these examples are shown.

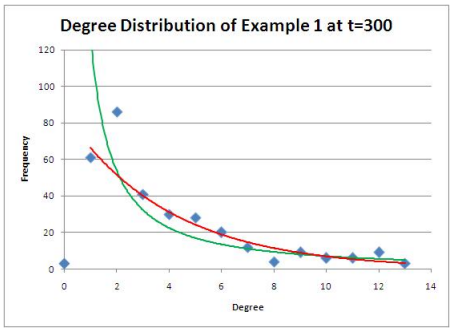

Figure 6: The degree distribution of a graph at $t = 300$ with a fitted power law (shown in green) and exponential (shown in red)

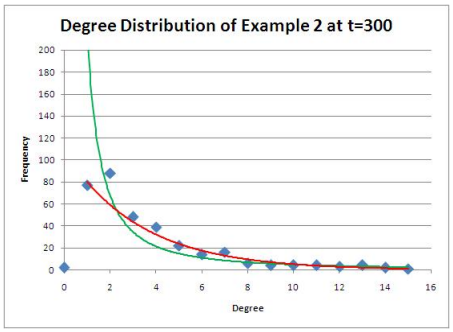

Figure 7: The degree distribution of a graph at $t = 300$ with a fitted power law (shown in green) and exponential (shown in red)

Disregarding the number of isolated vertices (vertices of degree zero), a power law and a decaying exponential were fitted to the data. For the power law, $R^2 = 0.794$ for the first example and $R^2 = 0.878$ for the second example. The exponential has $R^2 = 0.810$ for the first example and $R^2 = 0.947$ for the second example. In both cases, the decaying exponential is a closer fit. Recall the conclusion of Barabási and Albert. If a model incorporates growth without preferential attachment, the degree sequence follows a decaying exponential [1]. Therefore, we should not be surprised that our model approximately follows a decaying exponential rather than a power law.

9 Conclusions and Future Work

The simulations are more or less consistent with our expected results for order, size, and distribution of degrees. Yet despite all of the examples and results that have been presented, there are still many aspects that should be explored further. We would like to incorporate some different ideas into our model and examine some additional properties.

9.1 Diameter

The *distance* between two vertices u and v is the shortest number of edges in a trail between u and v. The *diameter* of a graph is the largest distance between any two vertices in the graph.

Bollobás and Riordan were able to put bounds on the diameter of a graph constructed with their model [3]. We are interested in doing something similar for our model. Due to time constraints, we were unable to explore the diameter in detail. However, this is a future direction for the project.

9.2 Add Preferential Attachment

The involvement of preferential attachment has been observed to change the expected order and size of a model. We now revisit the formulas for these properties in our model and inspect how they may be different when preferential attachment is included in the model.

Considering the expected order, preferential attachment will only affect which vertices a new vertex connects to, not whether or not a vertex is added. So at any time, the expected number of vertices in the graph only depends

on p_1, p_2, and p_5. Since these are parameters that will not be changed by the inclusion of preferential attachment, the expected order for a preferential attachment model should remain the same.

The expected size of the graph, however, will be altered. Let us consider the terms of $\tilde{M}$ separately. The first term, $t[p_1 + (2 + h_0)p_2 + p_3 - p_4]$, counts the expected number of edges that would be added only from which step is taken at a time t. Again, preferential attachment will only change where these edges are attached, not whether or not they are there. Therefore, this term of the expected size will be unaffected by preferential attachment. The second term, $[h_1(p_1 + p_2 + p_2h_0h_2) - p_5] \sum_{i=0}^{t-1}(2\tilde{M}(i)/N(i))$, counts the expected number of edges added to the neighborhood of a selected vertex. This term has two distinct parts. First, $[h_1(p_1 + p_2 + p_2h_0h_2) - p_5]$ calculates the chance that edges will be added or removed from a neighborhood and $\sum_{i=0}^{t-1}(2\tilde{M}(i)/N(i))$ calculates the average degree of a vertex at every previous time. The chance that edges will be added or removed from a neighborhood remains constant for every vertex and there is no reason for preferential attachment to change that. The average degree of a vertex, however, will be greatly changed. Since the vertices with many edges are more likely to gain additional edges, the expected degree of a vertex would no longer be $2\tilde{M}(t)/N(t)$. Thus, it is the only part of the expected size that will be affected by the addition of preferential attachment. Our conjecture is that this portion of the formula (and subsequently $\tilde{M}(t)$ as a whole) will increase due to preferential attachment.

With additional analysis, we would like to explore these conjectures. In that case, it would be interesting to revisit the distribution of degrees as well. Barabási and Albert showed that the addition of preferential attachment can change the degree sequence from a decaying exponential to a power law [1]. However, in Chung and Lu's duplication-deletion model and their mixed duplication model, the degrees followed a power law without using preferential attachment [6]. Since our model incorporates features of all three of these models, it would be interesting to see how the degree distribution changes with preferential attachment.

9.3 Non-traditional Preferential Attachment

Observe that in our model, the eight probability parameters are absolutely constant. When a single vertex connects to a selected vertex, the probability that it will add to the neighborhood of that selected vertex will always

be h_1, no matter how many other vertices are in that neighborhood. When preferential attachment is included, it will make vertices more likely to select a highly connected vertex. If h_1 is absolutely constant, the new vertex is just as likely to connect to the large neighborhood as it is to a smaller neighborhood. This is not what we see in a system like Facebook. If an existing profile has a very small number of friends, a new friend is more likely to know all of the friends of that profile, whereas if the existing profile has a very large number of friends, a new friend is not as likely to know them all. Therefore, the probability that a vertex connects to the neighborhood of the selected vertex should decrease with the degree of the selected vertex.

This idea is very similar to preferential attachment, but it is distinct because it is not concerned with just where the edges are added, but rather whether or not they will be added at all. Hence it is a "non-traditional" preferential attachment.

On a related note, we are also interested in modifying the meaning of p_5 so that it will depend on the degree of a vertex. In this case, the probability of deleting a vertex would decrease with the degree of a vertex. That is, vertices with low degrees would be more likely to be selected for deletion. This idea is also makes sense in a Facebook system. A profile that has very few friends is more likely to be deleted than a profile that has many friends.

Including these ideas in our model may drastically change some of the global properties. Continuing the development of our model will be very interesting. Hopefully it will answer some of the questions that have been raised.

References

[1] Barabási, Albert-László; Albert, Réka. Emergence of scaling in random networks. *Science* 286 (1999), no. 5439, 509-512.

[2] Barabási, Albert-László Linked: the new science of networks. Perseus Publishing, Cambridge, Massachusetts, 2002. 280 pp. ISBN: 0-7382-0667-9

[3] Bollobás, Béla; Riordan, Oliver. The diameter of a scale-free random graph. *Combinatorica* 24 (2004), no. 1, 5-34.

[4] Bollobás, Béla; Riordan, Oliver; Spencer, Joel; Tusnády, Gábor. The degree sequence of a scale-free random graph process. *Random Structures Algorithms* 18 (2001), no. 3, 279-290.

[5] Chartrand, Gary. *Introductory Graph Theory*. Corrected reprint of the 1977 original. Dover Publications, Inc., New York, 1985. xii+294 pp. ISBN: 0-486-24775-905-01

[6] Chung, Fan; Lu, Linyuan. *Complex Graphs and Networks*. CBMS Regional Conference Series in Mathematics, 107. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 2006. viii+264 pp. ISBN: 978-0-8218-3657-6; 0-8218-3657-9