Random evolution in massive graphs

William Aiello [*]   Fan Chung [†]   Linyuan Lu [‡]

Abstract

Many massive graphs (such as WWW graphs and Call graphs) share certain universal characteristics which can be described by so-called the "power law". In this paper, we will first briefly survey the history and previous work on power law graphs. Then we will give four evolution models for generating power law graphs by adding one node/edge at a time. We will show that for any given edge density and desired distributions for in-degrees and out-degrees (not necessarily the same, but adhered to certain general conditions), the resulting graph will almost surely satisfy the power law and the in/out-degree conditions. We will show that our most general directed and undirected models include nearly all known models as special cases. In addition, we consider another crucial aspects of massive graphs that is called "scale-free" in the sense that the frequency of sampling (w.r.t. the growth rate) is independent of the parameter of the resulting power law graphs. We will show that our evolution models generate scale-free power law graphs[1].

1 Introduction

The number of Internet host as of January 2000 topped 70 million and is estimated to be growing at 63% per year [39]. The number of web pages indexed by large search engines now exceeds 500 million and it is estimated that over 4,000 web sites are created everyday. Is it possible to determine

[*] AT&T Labs, Florham Park, New Jersey.

[†] University of California, San Diego

[‡] University of California, San Diego

[1] An extended abstract appeared in *The 42th Annual Symposium on Foundation of Computer Sciences*, (2001), 510-519. This paper will appear in *Handbook on Massive Data Sets*, (Eds. J. Abello et al.).

simple structural properties for such massive and dynamic graphs as the Internet and the World Wide Web? For example, are these graphs connected? If not, what is the size and diameter of the largest component? Are there interesting structural properties which govern or influence the development and use of these physical and virtual networks?

Of course, answering these questions exactly is quite likely not possible. However, in many other areas of the physical, biological, and social sciences and in engineering where the size and dynamic nature of the data sets similarly do not allow for exact answers, progress in understanding has nonetheless been achieved through an iterative interplay between experimental data and modeling, where both the data and the modeling often have a random or statistical basis. Such an interplay is in its early stages for the study of several massive, dynamic graphs such as the World Wide Web. The starting point of this interplay began when several groups independently made an important observation: the degree distributions of several different massive graphs, including the WWW graph, follow a power law [7, 8, 24, 10]. In a power law degree distribution, the fraction of nodes with degree d is proportional to $1/d^{\alpha}$ for some constant $\alpha \geq 0$. In this paper we present and analyze a general random graph evolution model which yields graphs with power law degree distributions. Below we will first review the empirical findings for graphs with power law degree distributions followed by an overview of previous modeling work for such graphs. Then we will discuss the models and results presented in this paper. In particular, we will examine the three important aspects of power law graphs, (1) analyzing the evolution of graphs, (2) the asymmetry of in-degrees and out-degrees, (3) the "scale invariance" of power law graphs.

2 History of power law graphs

2.1 Early history

The history of power laws can be traced back to statistical analysis in a variety of fields, including linguistics, academic citation, physical sciences, or even in nature or economy. In 1926, Lotka [27] plotted the distribution of authors in the decennial index of Chemical Abstracts (1907-1916), and he found that the number of authors is inversely proportional to the square of the number of papers published by those authors (which is often called *Lotka's law* or *inverse square law* and *Yule's law* [48]). Zipf [50] observed that the frequency of English words follows a power law function. That is, the word frequency that has rank i among all word frequencies is proportional to $1/i^a$ where a is close to 1. This is called *Zipf's law* or *Zipf's distribution.* As Simon [43] noted in an influential paper in 1957, this distribution is also common to various phenomena, such as word frequencies in large samples of prose, city sizes and income distributions. There has been a large number of research papers on power laws in natural language [33, 40, 44], bibliometrics [15, 18, 20, 23, 42] social sciences [36, 21, 30] and nature [32, 31, 41].

2.2 Empirical power laws

Power laws in massive graphs have recently been reported in a variety of context. In 1999, Kumar et al. [24] reported that a web crawl of a pruned data set from 1997 containing about 40 million pages revealed that the in-degree and out-degree distributions of the web followed a power law. Albert and Barabasi [7, 8] independently reported the same phenomenon on the approximately 325 thousand node `nd.edu` subset of the web. Both reported a power of approximately 2.1 for the in-degree power law and 2.7

for the out-degree (although the degree sequence for the out-degree deviates from the power law for small degree). More recently, these figures have been confirmed for a Web crawl of approximately 200 million nodes [10]. Thus, the power law fit of the degree distribution of the Web appears to be remarkably stable over time and scale.

Faloutsos et al. [19] have also observed a power law for the degree distribution of the Internet network. They reported that the distribution of the out-degree for the interdomain routing tables fits a power law with a power of approximately 2.2 and that this power remained the same over several different snapshots of the network. At the router level the out-degree distribution for a single snapshot in 1995 followed a power law with a power of approximately 2.6.

In addition to the Web graph and the Internet graph, several other massive graphs exhibit a power law for the degree distribution. The graph derived from telephone calls during a period of time over one or more carriers' networks is called a call graph. Using data collected by Abello et al. [1], Aiello et al. [3] observe that their call graphs are power law graphs. Both the in-degrees and the out-degrees have a power of 2.1. The graphs derived from the U.S. power grid and from the co-stars graph of actors (where there is an edge between two actors if they have appeared together in a movie) also obey a power law [7] Thus, a power law fit for the degree distribution appears to be a ubiquitous and robust property for many massive real-world graphs.

2.3 Modeling Power Law Graphs

As discussed above, many of the graphs above are so large and dynamic that answering simple structural questions exactly by empirical means is very difficult or infeasible. It is important, therefore, to develop models which match empirically observed behavior and yet are themselves amenable to structural analysis. Good models often guide further empirical analysis which often subsequently requires the models to be refined, and so on.

To begin our discussion of modeling power law graphs, first note that the standard random graph models, $\mathcal{G}(n, p)$, $\mathcal{G}(n, |E|)$, and $\tilde{\mathcal{G}}^n$, will not suffice (see, for example, [6]). In these models, the choice of edges have a high degree of independence. Hence, the distribution of degrees decays exponentially from the expected or average degree.

In order for a power law degree distribution to emerge, the choice of edges must be correlated. To achieve this correlation, two basic approaches have been taken thus far. We will review them in turn. The first basic approach is exemplified in Aiello et al. [3]. They do not attempt to explain how graphs with a power law degree distribution arise. Rather, they focus on classes of graphs with a power law degree distribution and they derive the structures and properties (such as connected components [3], diameters [28], etc.) as a function of the power. Chung and Lu [12, 13] further extend the analysis to random graphs with arbitrary degree distribution. Newman et al. [38] take a similar approach but use different methods of analysis. Other remarkable works in this direction include Molloy and Reed [34, 35], and Łuczak [29]. Certain questions are likely to prove more amenable to analysis using the later approach than the former and vice versa. Thus, the two approaches are complementary.

The second approach to modeling power law graphs attempts to model the evolution of such graphs and the manner in which the power law degree distribution arises. We will briefly overview the history along the following three aspects of power law graphs, (1) the evolution of graphs, (2) the asymmetry of in-degrees and out-degrees, (3) the “scale-free” phenomenon.

2.3.1 The evolution of power law graphs

For example, in [7], Barabasi and Albert describe the following graph evolution process. They start with a small initial graph. At each time step they add a new node and an edge between the new node and each of m random nodes in the existing graph, where m is a parameter of the model. The random nodes are not chosen uniformly. Instead, the probability of picking a node is weighted according to its existing degree (the edges are assumed to be undirected). That is, if there are e_t edges at time t and node v has degree $\delta_{v,t}$ at time time t, then the probability of picking node v is $\delta_{v,t}/2e_t$. Using heuristic analysis (e.g., the analysis assumes that the discrete degree distribution is differentiable) they derive a power law for the degree distribution with a power of 3, regardless of m. Clearly, the fact that the power is 3 regardless of the parameter m is a drawback of the model. Moreover, it can easily be shown that all of edges (except, perhaps, those of the small initial graph) of a resulting graph can be decomposed into m disjoint forests (i.e., the graph has arboricity m). Presumably, most massive real-world graphs with power law degree distributions have a richer structure than this. As we will see, by inserting the appropriate parameters into our general model, our analysis does yield a degree distribution power law with power 3. A power law with power 3 for the degree distribution of this model was independently derived by Bollobás et al. [9].

The main intuition behind the development of a power law degree distribution for this model is as follows. Nodes which acquire a relatively large degree early on in the process have an “advantage” and continue to accumulate added degree because of the preferential selection of nodes with high degree. Barabasi and Albert show that if the preferential selection of high degree nodes is replaced by a uniform selection of nodes then the power law behavior of the degree distribution does not result. Moreover, if the number of nodes is fixed, as opposed to constantly increasing, then the power law degree distribution again fails to occur.

Kumar et al. also describe a random graph evolution process [25]. Unlike that of [7], their random graphs are directed. Their model has the advantage that the power in the power-law is a function of a parameter of the model. Their model is as follows. A node and an edge are added at every time step. With probability $1-\alpha$, a directed self-loop is added to the new node. With probability α, an edge is added from the new node to a randomly selected node. The node is selected in proportion to its current in-degrees. That is, since there are t edges at time t, the probability of picking node v at time t is $\delta^{\text{in}}_{v,t}/t$ where $\delta^{\text{in}}_{v,t}$ is the in-degree of v at time t. They analyze this evolution process with a heuristic analysis and they derive a power law for the degree distribution with a power of $1/\alpha$. As we will see, this model is a special case of our general model for which our analysis yields a power of $1+1/\alpha$. The above model has a similar drawback as that of [7]: the resulting random graph is a tree.

2.3.2 Asymmetry of in-degrees and out-degrees

Kumar et el. [25] provide a general model which they call the (α, β) model which has the advantage that the in-degree and the out-degree both follow a power law. The powers in the power law for the in-degree and out-degree need not be the same; they can be controlled independently by α and β. As before a node and an edge are added at every time step. Let w_t be the node added at step t. At each time step, two nodes are chosen from the existing graph. Node u is selected according to its out degree, i.e., the probability that u is chosen is $\delta_{u,t}^{\text{out}}/t$. Node v is selected according to its in degree, i.e., the probability that v is chosen is $\delta_{v,t}^{\text{in}}/t$. Then two coins are tossed. The "origin" coin is "u" with probability α and "w_t" with probability $1-\alpha$. The "destination" coin is "v" with probability β and "w_t" with probability $1-\beta$. The new edge is added from the outcome of the origin coin to the outcome of the destination coin. That is, an edge is added from: u to v with probability $\alpha\beta$; from u to w_t with probability $\alpha(1-\beta)$; from w_t to v with probability $(1-\alpha)\beta$, and from w_t to w_t with probability $(1-\alpha)(1-\beta)$. They claim an out-degree power law with a power of $1/\alpha$ and an in-degree power law with a power of $1/\beta$. (As with their first model, the (α, β) model is a special case of our model. Our analysis yields power laws with powers $1 + 1/\alpha$ and $1 + 1/\beta$ for the out-degree and in-degree, respectively.)

While the above model allows for different powers laws for the in-degree and out-degree and yields graphs which do not have small arboricity, it has the following restrictive property. Suppose that at time step t, the origin and destination coins are w_t and v, respectively. In this case, w_t will have out-degree 1 and in-degree 0 at time $t + 1$. Hence, w_t cannot be chosen as node v in time step $t + 1$ and thus its in-degree will be 0 at time $t + 2$. Continuing in this manner, w_t will always have in-degree 0. Thus, with high probability, a constant fraction (approximately $(1-\alpha)\beta$) of the nodes will have in-degree 0. Likewise, with high probability, a constant fraction (approximately $\alpha(1-\beta)$) of the nodes will have out-degree 0. While some real-world power law graphs may have this property, it is likely that some, e.g., the Web, do not, and a more general model would be desirable. Also note that this model is restricted to graphs with density 1 since one node and one edge are added at every time step.

Recently, Kumar at el. [26] proposed three evolution models — “linear growth copying”, “exponential growth copying”, and “linear growth variants”. The *Linear growth coping* model adds one new vertex with d out-links at a time. The destination of i-th out-link of the new vertex is either copied from the corresponding out-link of a “prototype” vertex (chosen randomly) or a random vertex. They showed that the in-degree sequence follows the power law. These models were designed explicitly to model the World Wide Web. Indeed, they show that their model has a large number of complete bipartite subgraphs, as has been observed in the WWW graph, whereas several other models, including that of [3], do not. This (and the linear growth variants model) has the similar drawback as the first model in [25]. The out-degree of every vertex is always a constant. Edges and vertices in the *exponential growth copying* model increases exponentially. This exponential growth copying model does not have the same drawback as the other two models have. However, it is not clear whether its out-degrees satisfy the power law distribution.

2.3.3 Scale-free property for power law graphs

Power-laws or heavy tailed distributions are often associated with self-similarity and scaling laws. Indeed, by comparing the web crawls of [7, 8] and [10, 24] we see that the same power law appears to govern various subgraphs of the web as well as the whole. However, while some subgraphs obey the same power law and appear to be self-similar, clearly, there exists subgraphs of the web which would not obey the power law (e.g., the subgraph defined by all nodes with out-degree 100). The natural problem is thus: formally define and analyze a scale-free property for power law graphs. While there may be several types of scaling behavior exhibited by power law graphs, to the best of our knowledge, we give the first such definition and show that our model exhibits this scale-free property.

3 Our Results

Below we will describe a sequence of graph evolution models. The first three, Models A, B, and C, are for directed graphs and are increasingly more general. The first two are primarily illustrative although they may have merits as models in their own right due to their parsimony. Model C incompasses all of the directed graph models above, except that of [26]. We also describe a fourth model, Model D, which is the natural analogue of Model C for undirected graphs.

Consider the following simple model which we call model A. At each time step, a new node is added with probability $1 - \alpha$. The node starts with in-weight 1 and out-weight 1. Whenever the node is the origin (destination) of a new edge, the out-weight (in-weight) is increased by 1. That is, the in-weight (out-weight) of a node u at time t is just $w^{\text{in}}_{u,t} = 1 + \delta^{\text{in}}_{u,t}$ ($w^{\text{out}}_{u,t} = 1 + \delta^{\text{out}}_{u,t}$). With probability α a random edge is added to the existing nodes. The origin (destination) of the new edge is chosen proportional to the current in-weights (out-weights) of the nodes. That is, u (v) is chosen as the origin (destination) of the new edge at time t with probability $w^{\text{out}}_{u,t}/t$ ($w^{\text{in}}_{v,t}/t$). Note the expected number of edges in the graph is αt and the expected number of nodes is $(1-\alpha)t$. Call the ratio of the former to the latter $\Delta = \alpha/(1-\alpha)$ as it is a measure of the density of the graph. As a corollary to our general result, we will show that this model yields a power law with power $2+1/\Delta$ for both the in-degrees and the out-degrees. Thus, this model allows for graphs of varying density. For this model we also derive the *joint* distribution for the in-degrees and out-degrees. We show that the number of nodes with in-degree i and out-degree j is proportional to $(i+j)^{3+1/\Delta}$.

Note that when an edge is added among existing nodes, the probabilities concerning which edge is added are functions of the current degree distribution. Thus, the probability distribution of the new degree distribution is a function of the current degree distribution. This is difficult to solve recursively since the current degree distribution, itself, has a probability distribution. However, this means that the expected value of the new degree distribution is a function of the current degree distribution. Moreover, as we will see, the change in the degree distribution from from step to step is bounded. Thus, we observe that the evolution of the degree distribution is a semi martingale where deviation from the expected value of the final degree distribution occurs with exponentially small tails. Due to linearity of expectation, we are able to solve for the expected value of the final degree distribution recursively. These recursive equations and their solutions are non-standard, to the best of our knowledge, and may be of independent interest.

One drawback of model A is that the density parameter Δ and the power in the power law cannot be controlled independently. They are both functions of the parameter α. Moreover, the in and out degree have the same power. A simple modification to model A yields model B which overcomes both drawbacks. When a new node is added with probability $1-\alpha$ at a time step, it will be given in-weight γ^{in} and out-weight γ^{out}. Thus, the in-weight (out-weight) of a node u at time t is just $w^{\text{in}}_{u,t} = \gamma^{\text{in}} + \delta^{\text{in}}_{u,t}$ ($w^{\text{out}}_{u,t} = \gamma^{\text{out}} + \delta^{\text{out}}_{u,t}$). As before, when an edge is added with probability α, the origin of the edge is chosen with probability proportional to the current out-weights and the destination is chosen with probability proportional to the current in-weights.

We will show that this graph evolution process yields graphs with power law degree distributions with powers $2+\gamma^{\text{in}}/\Delta$, and $2+\gamma^{\text{out}}/\Delta$ for the in-degrees and out-degrees, respectively. Note that the powers for the in-degrees and out-degrees and the density can all be controlled separately. This is the simplest model of which we are aware for which this is the case. Moreover, the model does not suffer from any of the other drawbacks mentioned above such as small arboricity or a constant fraction of nodes with no incoming edges.

While the above model may indeed be the simplest with which to model a real-world power law graph on the basis of measurements of the density of the graph and the powers for the in-degrees and out-degrees, it may not capture other features of the graph which are measurable. Hence, we would also like a more general model which, for example, would include the above model as well as that of [25]. Consider now model C. Suppose that at each time step four numbers $m^{e,e}, m^{n,e}, m^{e,n}, m^{n,n}$ are drawn according to some probability distribution. We assume that the four random variables are bounded. These four random variables need not be independent. In this time step $m^{e,e}$ edges are added between existing nodes in the graph. Of course, as before, the origin and destination of these edges are chosen independently according to the current out-degrees and in-degrees, respectively. Likewise, $m^{n,e}$ edges are added from the new node to existing nodes chosen independently according to the current in degrees. Likewise, $m^{e,n}$ edges are added from existing nodes (chosen independently according to the current out-degrees) to the new node. Finally, $m^{n,n}$ directed self loops are added to the new node. We will ignore nodes which are born with no indegree or outdegree (i.e., at the time step the node is born $m^{n,n} = m^{e,n} = m^{n,e} = 0$), or alternatively we will not include degree zero in the degree distribution. Of course, each of these random variables has a well-defined expectation which we denote $\mu^{e,e}, \mu^{n,e}, \mu^{e,n}, \mu^{n,n}$, respectively. We show that this general process still yields a power law degree distribution. We derive a power of $2 + (\mu^{n,n} + \mu^{n,e})/(\mu^{e,n} + \mu^{e,e})$ for the out-degree. Consider the rightmost ratio in this expression. By definition, the first element of a superscript refers to the origination of the random edges. Hence, the numerator of this ratio is the expected number of edges per step with the new new node as the origin and the denominator is the expected number of edges per step with an existing node as the origin. We also derive a power of $2+(\mu^{n,n}+\mu^{e,n})/(\mu^{n,e}+\mu^{e,e})$ for the in-degree. Analogously to the expression for outdegree, recall that the the second element of a superscript refers to the destination of the random edges. Hence, the numerator of this ratio is the expected number of edges per step with the new new node as the destination and the denominator is the expected number of edges per step with an existing node as the destination.

Note that the first, simple model of [25] has $\mu^{n,e} = \alpha, \mu^{n,n} = 1 - \alpha$ and $\mu^{e,e} = \mu^{e,n} = 0$. Substituting this into our result gives an in-degree power of $2 + (1 - \alpha)/\alpha = 1 + 1/\alpha$. The (α, β) model of [25] gives $\mu^{e,e} = \alpha\beta, \mu^{n,e} = (1 - \alpha)\beta, \mu^{e,n} = \alpha(1 - \beta), \mu^{n,n} = (1 - \alpha)(1 - \beta)$. Using our general results this gives an out-degree power of $1 + 1/\alpha$ and an in-degree power of $1 + /\beta$. Also note that our model A has $\mu^{e,e} = \alpha$, $\mu^{e,n} = \mu^{n,e} = 0$ and $\mu^{n,n} = 1 - \alpha$. This yields a power of $1 + 1/\alpha$, as claimed, for both the in- and out-degrees. Model C can easily be generalized to include the parameters of the initial weights of the new nodes given in Model B but we omit that here.

Finally, we also describe a general undirected model which we denote Model D. It is a natural variant of Model C. At each time step three numbers $(m^{e,e}, m^{n,e}, m^{n,n})$ are drawn according to some probability distribution. We assume that the three random variables are bounded. In this time step

$m^{e,e}$ undirected edges are added between existing nodes in the graph. The endpoints of these edges are chosen independently according to the current total degrees. Likewise, $m^{n,e}$ edges are added between the new node and existing nodes chosen independently according to the current total degrees. Finally, $m^{n,n}$ undirected self loops are added to the new node. We show that this undirected graph evolution process also yields a power law degree distribution. We derive a power of $2 + (2\mu^{n,n} + \mu^{n,e})/(\mu^{n,e} + 2\mu^{e,e})$. Note that model of Barabasi and Albert [7] has $\mu^{n,n} = \mu^{e,e} = 0$ and $\mu^{n,e} = m$. Substituting this into our general result gives a power of 3 which matches their heuristically derived bound. Note that the natural undirected version of model A has $\mu^{n,e} = 0$ and thus a power of $2 + \mu^{n,n}/\mu^{e,e} = 1 + 1/\alpha$. As with model C, initial weights can easily be incorporated into Model D.

We remark that our conditions for Model C and D are much weaker than the previous known models. For example, previous known models assume that the way in which edges are added are identical at each time. In our models, to analyze the asymptotic value of the expectation of the degree distribution, we only need to assume edges are added in an "asymptotically similar" way.

Scale Invariance The evolution of massive graphs can be viewed as a process of growing graphs by adding nodes and edges at a time. One way is to divide the time into almost equal units and combine all nodes born in the same unit time into one super-node. The bigger time unit one chooses, the smaller size of the result graph has. This procedure is similar to scaling maps in space. The property is called *scale-free*. A model is called *scale-free* if it generates the scale-free power graphs with high probability. In other words, an evolution model is time scale invariant if we change the time scale by any given factor and examine the scaled graph, then the original graph and the scaled graph should satisfy the power law with the same powers for the in-degrees and out-degrees. Suppose that a “unit” of time is scaled by a factor of c. In other words, we combine all nodes born in previous c-units into one super node. This has the same effect adding edges c-times in a large one unit. A detailed definition will be given below.

Briefly, we scale time in our model and then show that the degree distribution of Model C is invariant with respect to the time scaling. To begin the discussion, consider a Model C evolution process with parameters $\mu^{n,n}$, $\mu^{n,e}$, $\mu^{e,n}$, and $\mu^{e,e}$ and a bound B on the number of edges added per time step. Suppose the evolution process is run for T time steps and let G_T be the graph generated. Label nodes by the time step in which they are added to the graph. To scale this evolution process by a factor of σ, we begin by aggregating time steps into super steps of σ consecutive time steps. That is, super-step 1 consists of time steps 1 through σ, super-step 2 consists of times steps $\sigma + 1$ through 2σ, and so on (where we assume for convenience that σ divides T). The scaled graph $H_\sigma(G_T)$ is created from G_T as follows. A node in G_T with step label i is mapped to the node in $H_\sigma(G_T)$ with super step label $\lceil i/\sigma \rceil$. (If there is no node in G_T with time label in super step τ then no node is created in G_T^σ with label τ.) An edge in G_T from node i to node j gets mapped to an edge in $H_\sigma(G_T)$ from node $\lceil i/\sigma \rceil$ to node $\lceil j/\sigma \rceil$. The morphism H_σ on this evolution process of Model C defines a natural evolution process, which, strictly speaking, is not covered by Model C. Nonetheless, we will show that this evolution process has the same power law asymptotically as a Model C evolution process with parameters $\mu'^{n,n} = \sigma\mu^{n,n}$, $\mu'^{n,e} = \sigma\mu^{n,e}$ $\mu'^{e,n} = \sigma\mu^{e,n}$ $\mu'^{e,e} = \sigma\mu^{e,e}$ and size bound σB. Given our general results on Model C, the latter Model C process has the same power law as the first Model C process (e.g., the power for the out-degree is $2 + (\mu^{n,n} + \mu^{n,e})/(\mu^{e,n} + \mu^{e,e})$) and therefore the time scaled process defined by the morphism H_σ has the same power law as the first Model C process. Thus, the power law degree distribution of a Model C evolution process is invariant with respect to the time scaling defined above.

The rest of the paper is organized as follows. In section 4, we will define Models A,B,C,D, and state our Theorems (Theorems 1,2,3,4) on the power law degree distribution of these models. We also state the scale-free property of these models (Theorem 5). In section 5, we prove Theorem 1 while Theorem 3 and 5 are proved in Section 6. The proofs of Theorems 2 and 4 are omitted.

Variations of the model We have also considered a variant of our model which can be described as follows. As in model C, a four-tuple of random variables $(m^{e,e}, m^{n,e}, m^{e,n}, m^{n,n})$ is sampled every time step. However, rather than adding $m^{e,e}$ edges to the existing graph by choosing a separate origin and destination (according to degree) for each edge, we choose a single origin according to degree and choose a separate destination according to degree for each edge. The recurrence for the expected out-degree distribution is slightly different than that for model C but nonetheless the asymptotic power of the power law is the same: $2 + (\mu^{n,n} + \mu^{n,e})/(\mu^{e,n} + \mu^{e,e})$. And again we see that the ratio can be described as the expected number of edges with the new node as the origin over the expected number of edges with an existing node as the origin. More generally, several nodes can be choosen according to degree and from each such nodes several outgoing edges can be added in the usual degree-biased way. The important quantity is $\mu^{e,e}$. Analogously, the destination can be a single node choosen according to degree and the origins for all $m^{e,e}$ edges can be choosen according to degree. In this case, the outdegree is identical to that of model C and the indegree is asymptotically the same. Similar variants apply to the undirected model.

A recent paper has studied a similar model to that above. First, consider a simplified model of the undirected graph model variant above. Either a new node is born with a random number edges to the existing graph ($m^{n,n} = m^{e,e} = 0$ and $m^{n,e} \geq 0$) with probability $1 - \alpha$, or edges are added between a single node in the graph and a random number of other nodes in the graph ($m^{n,n} = m^{n,e} = 0$ and $m^{e,e} \geq 0$) with probability α. If $\tilde{\mu}^{n,e}$ is the expected value of $m^{n,e}$ in the former case and $\tilde{\mu}^{e,e}$ is the expected value of $m^{e,e}$ in the later case then the power law degree distribution has power $2 + (1 - \alpha)\tilde{\mu}^{n,e}/((1 - \alpha)\tilde{\mu}^{n,e} + 2\alpha\tilde{\mu}^{e,e})$. A directed version of this model is analogous.

The model of Cooper and Frieze [14] effectively takes this simplified model as a starting point and then allows all choices of nodes for origins or destinations in the existing graph to be either sampled according to degree or sampled uniformly. The choice is determined by a biased coin. When the sampling of nodes is according to degree with probability one, the models of [14] reduce to those above. Cooper and Frieze also argue that their model produces asymptotically the same degree sequence as that of [26].

4 A General Graph Evolution Process

4.1 Notations and definitions

In [3], a random graph model was proposed which can be viewed as a special case of sparse random graphs with given degree sequences. This model involves two parameters, called logsize, denoted by α and log-log growth

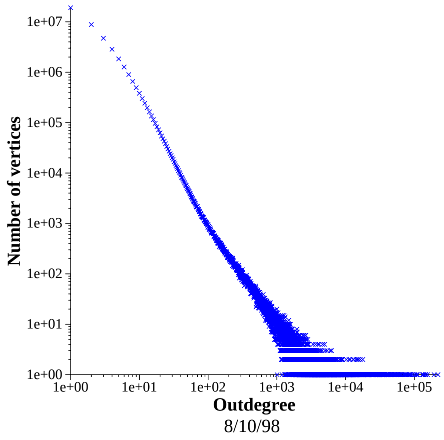

Figure 1: *The in-degree of call graph*

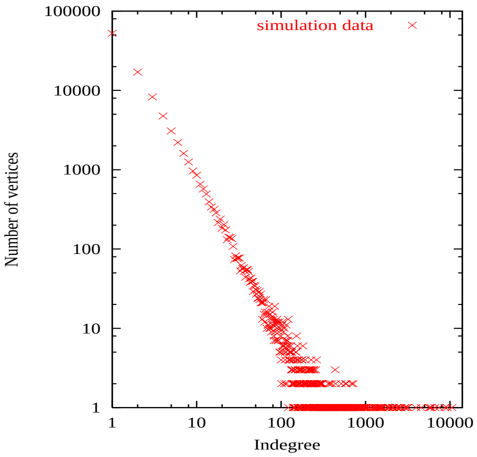

Figure 2: *Simulation using model A*

rate, denoted by β. A random graph under consideration has the following degree distribution: Suppose there are y vertices of degree x where x and y satisfy

$$\log y = \alpha - \beta \log x$$

Such graphs are called *power law graphs* with parameters (α, β). As it turns out that the parameters capture some universal characteristics of massive graphs. Furthermore, from these parameters, various properties of the graph can be derived. For example, for certain ranges of the parameters, we can compute the expected distribution of the sizes of the connected components which almost surely occur with high probability [3].

For a directed graph, the in-degree and out-degree sequence may follow power laws with different powers, as shown in massive graphs such as the Web graphs.

A robust way to generate a power law graph is to consider a random process, which grows the graph by adding one node and some edges at a time. Now we will give the definition of four models.

4.1.1 Model A

Model A is the basic model which the subsequent models rely upon. It starts with no node and no edge at time 0. At time 1, a node with in-weight 1 and out-weight 1 is added. At time $t+1$, with probability $1-\alpha$ a new node with in-weight 1 and out-weight 1 is added. With probability α a new directed edge uv is added to the existing nodes. Here the origin u is chosen with probability proportional to the current out-weight $w_{u,t}^{out} \overset{def}{=} 1 + \delta_{u,t}^{out}$ and the destination v is chosen with probability proportional to the current in-weight $w_{v,t}^{in} \overset{def}{=} 1 + \delta_{v,t}^{in}$. We note that $\delta_{u,t}^{out}$ and $\delta_{v,t}^{in}$ denote the out-degree of u and the in-degree of v at time t, respectively.

The total in-weight (out-weight) of graph in model A increases by 1 at a time. At time t, both total in-weight and total out-weight are exactly t. So the probability that a new edge is added onto two particular nodes u and v is exactly

$$\alpha \frac{(1+\delta_{u,t}^{out})(1+\delta_{v,t}^{in})}{t^2}.$$

The complete analysis will be given completely in next section.

4.1.2 Model B

Model B is a slight improvement of Model A. Two additional positive constant γ^{in} and γ^{out} are introduced. Different powers can be generated for in-degrees and out-degrees. In addition, the edge density can be independently controlled.

Model B starts with no node and no edge at time 0. At time 1, a node with in-weight γ^{in} and out-weight γ^{out} is added. At time $t+1$, with probability $1-\alpha$ a new node with in-weight γ^{in} and out-weight γ^{out} is added.

With probability α a new directed edge uv is added to the existing nodes. Here the origin u (destination v) is chosen proportional to the current out-weight $w_{u,t}^{out} \overset{def}{=} \gamma^{out} + \delta_{u,t}^{out}$ while the current in-weight is $w_{v,t}^{in} \overset{def}{=} \gamma^{in} + \delta_{v,t}^{in}$. Here $\delta_{u,t}^{out}$ is the out-degree of u and $\delta_{v,t}^{in}$ is the in-degree of v at time t, respectively.

In model B, at time t the total in-weight w_t^{in} and the out-weight w_t^{out}) of the graph are random variables. The probability that a new edge is added onto two particular nodes u and v is

$$\alpha \frac{(\gamma^{out} + \delta_{u,t}^{out})(\gamma^{in} + \delta_{v,t}^{in})}{w_t^{in} w_t^{out}}.$$

4.1.3 Model C

Now we consider Model C, this is a general model with four specified types of edges to be added.

Assume that the random process of model C starts at time t_0. At $t = t_0$, we start with an initial directed graph with some vertices and edges. At step $t > t_0$, a new vertex is added and four numbers $m^{e,e}, m^{n,e}, m^{e,n}, m^{n,n}$ are drawn according to some probability distribution. (Indeed, any bounded distribution is allowed here. It can even be a function of time t as long as the limit distribution exists as t approaches infinity.) We assume that the four random variables are bounded. Then we proceed as follows:

- Add $m^{e,e}$ edges randomly. The origins are chosen with the probability proportional to the current out-degree and the destinations are chosen proportional to the current in-degree.

- Add $m^{e,n}$ edges into the new vertex randomly. The origins are chosen with the probability proportional to the current out-degree and the destinations are the new vertex.

- Add $m^{n,e}$ edges from the new vertex randomly. The destinations are chosen with the probability proportional to the current in-degree and the origins are the new vertex.

- Add $m^{n,n}$ loops to the new vertex.

Each of these random variables has a well-defined expectation which we denote by $\mu^{e,e}, \mu^{n,e}, \mu^{e,n}, \mu^{n,n}$, respectively. We will show that this general process still yields power law degree distributions and the powers are simple rational functions of $\mu^{e,e}, \mu^{n,e}, \mu^{e,n}, \mu^{n,n}$.

4.1.4 Model D

Model A, B and C are all power law models for directed graphs. Here we describe a general undirected model which we denote by Model D. It is a natural variant of Model C.

We assume that the random process of model C starts at time t_0. At $t = t_0$, we start with an initial undirected graph with some vertices and edges. At step $t > t_0$, a new vertex is added and three numbers $m^{e,e}, m^{n,e}, m^{n,n}$ are drawn according to some probability distribution. We assume that the three random variables are bounded. Then we proceed as follows:

- Add $m^{e,e}$ edges randomly. The vertices are chosen with the probability proportional to the current degree.

- Add $m^{e,n}$ edges randomly. One vertex of each edge must be the new vertex. The other one is chosen with the probability proportional to the current degree.

- Add $m^{n,n}$ loops to the new vertex.

4.1.5 General notations

For all (directed) graph models A, B, C, D, we denote n_t to be the number of vertices at time t. Let e_t be the number of edges at time t.

For graph models A, B, C, let $d_{i,t}^{in}$ and $d_{j,t}^{out}$ denote the random variables as the number of vertices with in-degree i and out-degree j, respectively. Let $d_{i,j,t}^{joint}$ be the random variable as the number of vertices with in-degree i and out-degree j.

For (undirected) graph model D, let $d_{i,t}$ denote the random variable as the number of vertices with degree i.

4.2 Results and applications

We first state the theorems that will be proved in latter sections.

Theorem 1 *For model A, the distribution of in-degree and out-degree sequences follow the power law distribution with power $1 + \frac{1}{\alpha}$. The joint distribution of in-degree and out-degree sequence follows the power law distribution with power $2 + \frac{1}{\alpha}$. More precisely, we have*

$$Pr(|d_{i,j,t}^{joint} - a_{i,j}t| > \lambda\sqrt{t} + 2) < e^{-\lambda^2/8},$$

$$Pr(|d_{i,t}^{in} - b_i t| > \lambda\sqrt{t} + 2) < e^{-\lambda^2/2},$$

$$Pr(|d_{j,t}^{out} - c_j t| > \lambda\sqrt{t} + 2) < e^{-\lambda^2/2}.$$

where $a_{i,j}, b_i, c_j$ satisfy

$$\begin{aligned} a_{i,j} &= (1-\alpha)\frac{(i+j-2)!\alpha^{i+j-2}}{\prod_{l=2}^{i+j}(1+l\alpha)} = \frac{(\frac{1}{\alpha}-1)\Gamma(\frac{1}{\alpha}+2)}{(i+j)^{\frac{1}{\alpha}+2}} + o_{i+j}(1) \\ b_i &= (1-\alpha)\frac{(i-1)!\alpha^{i-1}}{\prod_{l=1}^{i}(1+l\alpha)} = \frac{(\frac{1}{\alpha}-1)\Gamma(\frac{1}{\alpha}+1)}{i^{\frac{1}{\alpha}+1}} + o_i(1) \\ c_j &= (1-\alpha)\frac{(j-1)!\alpha^{j-1}}{\prod_{l=1}^{j}(1+l\alpha)} = \frac{(\frac{1}{\alpha}-1)\Gamma(\frac{1}{\alpha}+1)}{j^{\frac{1}{\alpha}+1}} + o_j(1) \end{aligned}$$

For all i,j,t, the expected values $E(d_{i,j,t}^{joint}), E(d_{i,t}^{in})$ *and* $E(d_{j,t}^{out})$ *satisfy*

$$\begin{aligned} |E(d_{i,j,t}^{joint}) - a_{i,j}t| &< 2 \\ |E(d_{i,t}^{in}) - b_i t| &< 2 \\ |E(d_{j,t}^{out}) - c_j t| &< 2. \end{aligned}$$

Theorem 2 *For model B, the distribution of in-degree sequence follows the power law distribution with power* $2 + \frac{\gamma^{in}}{\Delta}$*, and the distribution of out-degree sequence follows the power law distribution with power* $2 + \frac{\gamma^{out}}{\Delta}$*. Here* $\Delta = \frac{\alpha}{1-\alpha}$ *is the asymptotic edge density. More precisely, we have*

$$Pr(|d_{i,t}^{in} - b_i' t| > 2\lambda\sqrt{t}) < e^{-\lambda^2/2},$$

$$Pr(|d_{j,t}^{out} - c_j' t| > 2\lambda\sqrt{t}) < e^{-\lambda^2/2}.$$

where b_i', c_j' *satisfy*

$$\begin{aligned} b_i' &= (1-\alpha)(\frac{1}{\gamma^{in}} + \frac{1}{\Delta}) \prod_{l=1}^{i+1} \frac{l - 2 + \gamma^{in}}{l + \frac{\gamma^{in}}{\alpha}} \\ &= (1-\alpha)(\frac{1}{\gamma^{in}} + \frac{1}{\Delta}) \frac{\Gamma(\frac{\gamma^{in}}{\alpha} + 1)}{\Gamma(\gamma^{in} - 1)} \frac{1}{i^{\frac{\gamma^{in}}{\Delta} + 2}} + o_i(1) \\ c_j' &= (1-\alpha)(\frac{1}{\gamma^{out}} + \frac{1}{\Delta}) \prod_{l=1}^{j+1} \frac{l - 2 + \gamma^{out}}{l + \frac{\gamma^{out}}{\alpha}} \\ &= (1-\alpha)(\frac{1}{\gamma^{out}} + \frac{1}{\Delta}) \frac{\Gamma(\frac{\gamma^{out}}{\alpha} + 1)}{\Gamma(\gamma^{out} - 1)} \frac{1}{j^{\frac{\gamma^{out}}{\Delta} + 2}} + o_j(1) \end{aligned}$$

Theorem 3 *For model C, almost surely the out-degree sequence follows the power law distribution with the power* $2 + \frac{\mu^{n,n} + \mu^{n,e}}{\mu^{e,n} + \mu^{e,e}}$ *where* μ*'s are as defined in 2.1.3.) Almost surely the in-degree sequence follows the power law distribution with the power* $2 + \frac{\mu^{n,n} + \mu^{e,n}}{\mu^{n,e} + \mu^{e,e}}$*. More precisely, we have*

$$Pr(|d_{i,t}^{in} - b_i'' t| > 2M\lambda\sqrt{t}) < e^{-\lambda^2/2},$$

$$Pr(|d^{out}_{j,t} - c''_j t| > 2M\lambda\sqrt{t}) < e^{-\lambda^2/2}.$$

where b''_i, c''_j *satisfy*

$$b''_i = \frac{b''}{i^{2+\frac{\mu^{n,n}+\mu^{e,n}}{\mu^{n,e}+\mu^{e,e}}}} + o_i(1),$$

$$c''_j = \frac{c''}{j^{2+\frac{\mu^{n,n}+\mu^{e,n}}{\mu^{n,e}+\mu^{e,e}}}} + o_j(1).$$

Here b'', c'', M *are constants determined by the joint distribution of* $m^{e,e}$*,* $m^{n,e}$*,* $m^{e,n}$*,* $m^{n,n}$ *of this model, but independent of* i *and* t*. (See the proof in section 4 for definitions of* b'', c'', M*.)*

Theorem 4 *For model D, almost surely the degree sequence follows the power law distribution with the power* $2 + \frac{2\mu^{n,n}+\mu^{n,e}}{\mu^{n,e}+2\mu^{e,e}}$*. More precisely, we have*

$$Pr(|d^{in}_{i,t} - a'_i t| > 2M'\lambda\sqrt{t}) < e^{-\lambda^2/2},$$

where a'_i *satisfies*

$$a'_i = \frac{a'}{i^{2+\frac{2\mu^{n,n}+\mu^{n,e}}{\mu^{n,e}+2\mu^{e,e}}}} + o_i(1).$$

Here a', M' *are constants determined by distribution of* $(m^{e,e}, m^{n,e}, m^{n,n})$ *of this model, but independent of* i *and* t*.*

Theorem 3 has an important application on “Scale-free” property.

Theorem 5 *Model A, B, C, D are scale-free. Especially almost all previous models [7, 8, 24, 25] are scale-free.*

Remarks: Theorem 1 and 2 hold for all ranges of i, j, t. Theorems 3 and 4 hold for $t \geq t_0$, where t_0 depends on the initial graphs and the asymptotic behavior of the variables involved in the evolution process. In general, $d^{in}_{i,t}$ and $d^{out}_{j,t}$ concentrate on their expected values within an interval of length

$t^{1/2+\epsilon}$, for any $\epsilon > 0$. We note that the desirable range of i (or j) for Theorems 1-4 is $i \ll t^{1/(2p)}$, where p is the power in the power law model as stated in Theorems 1-4.

5 Proof of theorem 1

For models A,B,C,D, we denote $\mathcal{G}_t$ the probability space associated to each graph G_t at time t. As t increases, $\mathcal{G}_t$ can be defined recursively. For each t, let τ_t be a random variable of $\mathcal{G}_t$.

$\{\tau_t\}$ is said to satisfy the *c-Lipschitz condition.* if

$$|\tau_{t+1}(H_{t+1}) - \tau_t(H_t)| \leq c$$

whenever H_{t+1} is obtained from H_t by adding some edges or some vertices at time $t+1$.

This concept is very similar to the vertex or edge Lipschitz condition in classical random graph theory (see [5]). We will use the following fact which is from the standard martingale theory.

Lemma 1 *If τ satisfies the c-Lipschitz condition, then we have for every $\lambda > 0$*

$$Pr[|\tau_t - E(\tau_t)| > \lambda\sqrt{t}] < 2e^{-\frac{\lambda^2}{2c^2}}$$

In particular, τ_t is almost surely very close to its expected value $E(\tau_t)$ with an error term $o(t^{\frac{1}{2}+\varepsilon})$ for any $\varepsilon > 0$, as t approaches infinity.

Proof of Theorem 1: Both $\{d^{in}_{i,t}\}$ and $\{d^{out}_{j,t}\}$ satisfy 1-Lipschitz condition. $\{d^{joint}_{i,j,t}\}$ satisfies 2-Lipschitz condition. By Lemma 1, it is enough to compute the corresponding expected values. Here we compute $E(d^{joint}_{i,j,t})$ in detail.

At time 0, there is nothing in graph. At time 1, a node with a loop is added. So we have

$$d_{1,1,1}^{joint} = 1 \text{ and } d_{i,j,1}^{joint} = 0 \text{ for } i > 1 \text{ or } j > 1$$

$i = 1, j = 1$ is special. For $t \geq 1$, we have

$$d_{1,1,t+1}^{joint} = \begin{cases} d_{1,1,t}^{joint} + 1 & \text{w.p. } 1 - \alpha \\ d_{1,1,t}^{joint} - 1 & \text{w.p. } \alpha(2\frac{d_{1,1,t}^{joint}}{t}(1 - \frac{d_{1,1,t}^{joint}}{t}) + \frac{d_{1,1,t}^{joint}}{t^2}) \\ d_{1,1,t}^{joint} - 2 & \text{w.p. } \alpha((\frac{d_{1,1,t}^{joint}}{t})^2 - \frac{d_{1,1,t}^{joint}}{t^2}) \\ d_{1,1,t}^{joint} & \text{otherwise} \end{cases}$$

In general, we have

$$d_{i,j,t+1}^{joint} = \begin{cases} d_{i,j,t}^{joint} + 2 & \text{w.p. } \frac{(i-1)d_{i-1,j,t}^{joint}}{t} \frac{(j-1)d_{i,j-1,t}^{joint}}{t} \\ d_{i,j,t}^{joint} + 1 & \text{w.p. } \alpha\frac{(i-1)d_{i-1,j,t}^{join}}{t}(1 - \frac{(j-1)d_{i,j-1,t}^{join}}{t}) \\ & +\alpha\frac{(j-1)d_{i,j-1,t}^{join}}{t}(1 - \frac{(i-1)d_{i-1,j,t}^{join}}{t}) + \alpha\frac{(i-1)(j-1)d_{i-1,j-1,t}^{join}}{t^2} \\ d_{i,j,t}^{joint} - 1 & \text{w.p. } \alpha\frac{id_{i,j,t}^{join}}{t}(1 - \frac{jd_{i,j,t}^{join}}{t}) + \alpha\frac{jd_{i,j,t}^{join}}{t}(1 - \frac{id_{i,j,t}^{join}}{t}) + \alpha\frac{ijd_{i,j,t}^{join}}{t^2} \\ d_{i,j,t}^{joint} - 2 & \text{w.p. } \alpha\frac{ij(d_{i,j,t}^{join})^2}{t^2} - \alpha\frac{ijd_{i,j,t}^{join}}{t^2} \\ d_{i,j,t}^{joint} & \text{otherwise} \end{cases}$$

Let $N_t = (d_{i,j,t}^{joint})_{all\ i,j}$ denote the degree distribution at time t. We have

$$E(d_{1,1,t+1}^{joint}|N_t) = d_{1,1,t}^{joint} + 1 - \alpha - \alpha(\frac{2}{t} - \frac{1}{t^2})d_{1,1,t}^{joint}$$

For $(i,j) \neq (1,1)$, similarly, we have

$$E(d_{i,j,t+1}^{joint}|N_t) = d_{i,j,t}^{joint} + \frac{\alpha}{t}((i-1)(1 - \frac{j}{t})d_{i-1,j,t}^{joint} +$$
$$(j-1)(1 - \frac{i}{t})d_{i,j-1,t}^{joint} - (i + j - \frac{ij}{t})d_{i,j,t}^{joint})$$

Hence we have the following recurrence formula:

$$E(d_{1,1,t+1}^{joint}) = E(d_{1,1,t}^{joint})(1 - \alpha(\frac{2}{t} - \frac{1}{t^2})) + 1 - \alpha$$

For $(i,j) \neq (1,1)$, we have

$$\begin{aligned} E(d^{joint}_{i,j,t+1}) &= E(d^{joint}_{i,j,t})(1-\alpha\frac{(i+j)}{t}+\alpha\frac{ij}{t^2}) \\ &\quad +\frac{(i-1)\alpha}{t}(1-\frac{j}{t})E(d^{joint}_{i-1,j,t}) \\ &\quad +\frac{(j-1)\alpha}{t}(1-\frac{i}{t})E(d^{joint}_{i,j-1,t}) \end{aligned}$$

To examine the asymptotic behavior of $E(d^{joint}_{i,j,t})$, we want to express

$$E(d^{joint}_{i,j,t}) = a_{i,j}t + c_{i,j,t},$$

where $c_{i,j,t} = o(t)$ is a lower order term. To choose an appropriate value for $a_{i,j}$, we substitute it into above recurrence formula and let t approach infinity. We obtain

$$a_{1,1} = \frac{1-\alpha}{1+2\alpha}$$

For $(i,j) \neq (1,1)$ we have

$$a_{i,j} = \alpha\frac{(i-1)a_{i-1,j}+(j-1)a_{i,j-1}}{1+(i+j)\alpha}$$

The solution to the above recurrence is the following:

$$\begin{aligned} a_{i,j} &= \frac{(1-\alpha)(i+j-2)!\alpha^{i+j-2}}{\prod_{k=2}^{i+j}(1+k\alpha)} \\ &= \frac{(\frac{1}{\alpha}-1)\Gamma(\frac{1}{\alpha}+2)}{(i+j)^{\frac{1}{\alpha}+2}} + o_{i+j}(1) \end{aligned}$$

for all i,j.

It suffices to establish an upper bound for $c_{i,j,t}$. In fact, we will show that $c_{i,j,t} \leq 2$. This will be proved by induction. When $i=j=1$, $c_{1,1,t}$ satisfies the following recurrence formula

$$c_{1,1,t+1} = c_{1,1,t}(1-\alpha(\frac{2}{t}-\frac{1}{t^2})) + \alpha\frac{1-\alpha}{1+2\alpha}\frac{1}{t}$$

Since $c_{1,1,1} = \frac{3\alpha}{1+2\alpha} < 2$, by induction on t, we have

$$c_{1,1,t+1} \leq 2(1-\alpha(\frac{2}{t}-\frac{1}{t^2})) + \alpha\frac{1-\alpha}{1+2\alpha}\frac{1}{t} \leq 2.$$

For $i \geq 2$ or $j \geq 2$, $c_{i,j,t}$'s satisfy the following recurrence formula:

$$\begin{aligned} c_{i,j,t+1} &= (1-\alpha\frac{it+jt-ij}{t^2})c_{i,j,t} + \frac{(i-1)\alpha}{t}(1-\frac{j}{t})c_{i-1,j,t} + \frac{(j-1)\alpha}{t} \\ & (1-\frac{i}{t})c_{i,j-1,t} + \frac{\alpha}{t}(ija_{i,j} - (i-1)ja_{i-1,j} - i(j-1)a_{i,j-1}) \end{aligned}$$

Now we use induction on i, j, t to show $|c_{i,j,t}| \leq 2$. By induction hypothesis, we assume that $|c_{i,j,t}| < 2, |c_{i-1,j,t}| < 2, |c_{i,j-1,t}| < 2$. Now we have

$$\begin{aligned} |c_{i,j,t+1}| &\leq 2(1-\alpha\frac{it+jt-ij}{t^2}) + \frac{(i-1)\alpha}{t}(1-\frac{j}{t})2 \\ &\quad + \frac{(j-1)\alpha}{t}(1-\frac{i}{t})2 + \frac{\alpha}{t}2 \\ &= 2 - \frac{2\alpha}{t}(1-\frac{1}{t}) - \frac{2\alpha(i-1)(j-1)}{t^2} \\ &\leq 2. \end{aligned}$$

Thus we finished the induction step. (Here we use the fact that $ija_{i,j} - (i-1)ja_{i-1,j} - i(j-1)a_{i,j-1} < \sum_{ij} ija_{ij} = 2$.)

The other two recurences can be proved analogously. Actually, b_i and c_j can be derived from $a_{i,j}$ by observing that

$$d^{in}_{i,t} = \sum_{j\geq 0} d^{joint}_{i,j,t} \qquad \text{and} \qquad d^{out}_{j,t} = \sum_{i\geq 0} d^{joint}_{i,j,t}.$$

□

The proof of Theorem 2 is similar and will be omitted. Next section, we will prove Theorems 3 and 5.

6 The proofs of Theorems 3 and 5

We first prove the following lemma.

Lemma 2 *If a sequence a_t satisfies the recursive formula*

$$a_{t+1} = (1 - \frac{b_t}{t})a_t + c_t \text{ for } t \geq t_0$$

where $\lim_{t\to\infty} b_t = b > 0$ *and* $\lim_{t\to\infty} c_t = c$ *exists. Then* $\lim_{t\to\infty} \frac{a_t}{t}$ *exists and*

$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}$$

Proof: Since $\lim_{t\to\infty} b_t = b > 0$, there exists a t_1 satisfying $b_t > 0$ for all $t > t_1$. Notice

$$\frac{c}{1+b}(t+1) = (1 - \frac{b}{t})\frac{c}{1+b}t + c.$$

We have

$$\begin{aligned} |a_{t+1} - \frac{c}{1+b}(t+1)| &= |(1 - \frac{b_t}{t})(a_t - \frac{c}{1+b}t) + (b_t - b)\frac{c}{1+b} + c_t - c| \\ &\leq |a_t - \frac{c}{1+b}t| + s_t \end{aligned}$$

where $s_t = |(b_t - b)\frac{c}{1+b} + c_t - c| \to 0$ as t approaches infinity.

Now we use this inequality recursively. We have

$$|a_t - \frac{c}{1+b}t| \leq |a_{t_1} - \frac{c}{1+b}t_1| + \sum_{k=t_1}^{t-1} s_k = o(t).$$

Hence the limit $\lim_{t\to\infty} \frac{a_t}{t}$ exists and $\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}$. □

We now proceed to prove Theorem 3.

Proof of Theorem 3: Only bounded number of edges are added at a time in model C. Let's denote this bound by M. Now both $d^{in}_{i,t}$ and $d^{out}_{j,t}$ satisfy M-Lipschitz condition. By Lemma 1, it is enough to show that following limits exist.

$$\lim_{t\to\infty} \frac{E(d^{in}_{i,t})}{t} = \frac{b''}{i^{2+\frac{\mu^{n,n}+\mu^{e,n}}{\mu^{n,e}+\mu^{e,e}}}} + o_i(1). \tag{1}$$

$$\lim_{t\to\infty} \frac{E(d^{out}_{j,t})}{t} = \frac{c''}{j^{2+\frac{\mu^{n,n}+\mu^{n,e}}{\mu^{e,n}+\mu^{e,e}}}} + o_j(1). \tag{2}$$

where b'', c'' are some constants independent of i, j.

We will prove the equation (1). The proof of (2) is similar and will be omitted.

We assume that at time t, with probability $p^t_{i'j'k'l'}$, $m^{e,e} = i', m^{n,e} = j', m^{e,n} = k', m^{n,n} = l'$. The probability that a vertex of in-degree $i - s$ becomes a vertex i is exactly

$$\binom{i'+j'}{s}(\frac{i-s}{e_t})^s(1-\frac{i-s}{e_t})^{i'+j'-s}. \tag{3}$$

The probability that a vertex of in-degree i becomes a vertex of in-degree $i + s$ is exactly

$$\binom{i'+j'}{s}(\frac{i}{e_t})^s(1-\frac{i}{e_t})^{i'+j'-s}. \tag{4}$$

The new vertex is a vertex of in-degree i is exactly

$$\sum_{k'+l'=i,i',j'} p^t_{i'j'k'l'},$$

which is assume to be well-behaved. So its limit as t approaches ∞ exists. We denote it by p_i.

By linearity of the conditional expectation, we have

$$\begin{aligned}
E(d^{in}_{i,t+1}|G_t) &= \sum_{s\geq 1} d^{in}_{i-s,t} \sum_{i',j',k'l'} \binom{i'+j'}{s}(\frac{i-s}{e_t})^s(1-\frac{i-s}{e_t})^{i'+j'-s} p^t_{i'j'k'l'} \\
&\quad - d^{in}_{i,t} \sum_{s\geq 1} \sum_{i',j',k'l'} \binom{i'+j'}{s}(\frac{i}{e_t})^s(1-\frac{i}{e_t})^{i'+j'-s} p^t_{i'j'k'l'} \\
&\quad + d^{in}_{i,t} + \sum_{k'+l'=i,i',j'} p^t_{i'j'k'l'} \\
&= d^{in}_{i,t}\left(1 - \frac{i}{e_t}\sum_{i',j',k'l'}(i'+j')p^t_{i'j'k'l'}(1+o(1))\right) \\
&\quad + d^{in}_{i-1,t}\frac{i-1}{e_t}\sum_{i',j',k'l'}(i'+j')p^t_{i'j'k'l'}(1+o(1)) + p^+_i(1+o(1))
\end{aligned}$$

$$\begin{aligned} &= d_{i,t}^{in}\left(1 - i\frac{\mu^{n,e} + \mu^{e,e} + o(1)}{(\mu^{n,e} + \mu^{e,e} + \mu^{e,n} + \mu^{n,n})t})\right) \\ &+ d_{i-1,t}^{in}(1 - i\frac{\mu^{n,e} + \mu^{e,e} + o(1)}{(\mu^{n,e} + \mu^{e,e} + \mu^{e,n} + \mu^{n,n})t}) + p_i^{+}(1 + o(1)) \end{aligned}$$

Since $m^{e,e}, m^{n,e}, m^{e,n}, m^{n,n}$ are bounded by M, we have $p_i = 0$, for $i > M$. We derive the following recurrence formula.

$$E(d_{i,t+1}^{in}) = E(d_{i,t}^{in})(1 - i\frac{\mu^{n,e} + \mu^{e,e} + o(1)}{(\mu^{n,e} + \mu^{e,e} + \mu^{e,n} + \mu^{n,n})t})$$

$$+E(d_{i-1,t}^{in})\frac{(1 + o(1))(i-1)(\mu^{n,e} + \mu^{e,e})}{(\mu^{n,e} + \mu^{e,e} + \mu^{e,n} + \mu^{n,n})t}) + o(1) \qquad (5)$$

for $i > M$.

By induction on i and Lemma 2, equation (5) implies $\lim_{t\to\infty} \frac{E(d_{i,t}^{in})}{t}$ exists. Let denote it by b_i''. b_i'' satisfies

$$\begin{aligned} b_i'' &= \frac{\frac{b_{i-1}''(i-1)(\mu^{n,e}+\mu^{e,e})}{\mu^{n,e}+\mu^{e,e}+\mu^{e,n}+\mu^{n,n}}}{1 + \frac{i(\mu^{n,e}+\mu^{e,e})}{(\mu^{n,e}+\mu^{e,e}+\mu^{e,n}+\mu^{n,n})}} \\ &= \frac{i-1}{i+1+\frac{\mu^{e,n}+\mu^{n,n}}{\mu^{n,e}+\mu^{e,e}}} b_{i-1}'' \end{aligned}$$

Hence, we have,

$$\begin{aligned} b_i'' &= \frac{i-1}{i+1+\frac{\mu^{e,n}+\mu^{n,n}}{\mu^{n,e}+\mu^{e,e}}} b_{i-1}'' \\ &= b_M'' \prod_{k=M+1}^{i} \frac{k-1}{k+1+\frac{\mu^{e,n}+\mu^{n,n}}{\mu^{n,e}+\mu^{e,e}}} \\ &= b_M'' \frac{(i-1)!\Gamma(M+2+\frac{\mu^{e,n}+\mu^{n,n}}{\mu^{n,e}+\mu^{e,e}})}{(M-1)!\Gamma(i+2+\frac{\mu^{e,n}+\mu^{n,n}}{\mu^{n,e}+\mu^{e,e}})} \\ &\approx \frac{b''}{i^{2+\frac{\mu^{e,n}+\mu^{n,n}}{\mu^{n,e}+\mu^{e,e}}}} \end{aligned}$$

where

$b'' = b''_M \frac{\Gamma(2M+2+\frac{\mu^{e,n}+\mu^{n,n}}{\mu^{n,e}+\mu^{e,e}})}{(2M-1)!}$ is a constant.

Equation (1) is proved.

By Lemma 1, almost surely the out-degree sequence of G_t satisfies the power law distribution with power $2 + \frac{\mu^{e,n}+\mu^{n,n}}{\mu^{n,e}+\mu^{e,e}}$.

Similarly, we can show almost surely the in-degree sequence of G_t satisfies the power law distribution with power $2 + \frac{\mu^{n,e}+\mu^{n,n}}{\mu^{e,n}+\mu^{e,e}}$. □

The proof of Theorem 4 is similar to this one and will be omitted.

Next, we will prove Theorem 5.

Proof of Theorem 5: Model A and previous models [7, 8, 24, 25] are the special cases of Model C. We will prove that Model C has the scale-free property. The proofs for Models A, B and D are similar and will be omitted.

We suppose that the evolution G_T is scaled by a factor of σ. (See section 1 for the definition.) The scaled evolution $H_\sigma(G_T)$ is not exactly covered by Model C. But it is naturally approximated by an evolution $G_{T'}$ of Model C with parameters $\mu'^{n,n} = \sigma\mu^{n,n}$, $\mu'^{n,e} = \sigma\mu^{n,e}$ $\mu'^{e,n} = \sigma\mu^{e,n}$ $\mu'^{e,e} = \sigma\mu^{e,e}$ and size bound σM. Given our general results on Model C, the latter Model C process has the same power law as the first Model C process (e.g., the power for the out-degrees is $2 + (\mu^{n,n} + \mu^{n,e})/(\mu^{e,n} + \mu^{e,e})$). Hence, it is enough to show that both the scaled evolution $H_\sigma(G_T)$ and the approximating evolution $G_{T'}$ have the same power for the out-degrees (and in-degrees).

The evolution $H_\sigma(G_T)$ only differs from $G_{T'}$ in the way of adding edges. At each time unit, edges are added simultaneously in $G_{T'}$ while some edges in $H_\sigma(G_T)$ are added simultaneously and some are added sequentially. By examining the proof of Theorem 3, we find that the probabilities given in equations (3) and (4) are different. However, the main terms (occurred at s=1) stay the same. From the proof of Theorem 3, we conclude that both evolutions give the same power for the out-degrees as well as for the in-degrees. □

7 Problems and remarks

In this paper, we use techniques in random graph theory to analyze power law graphs. The analysis of the evolution of power law graph are considerably harder than that for the earlier model of random graphs with given degree sequence [3, 34, 35] since nodes which acquire a relatively large degree early on in the process have an advantage and are substantially different from the nodes that are added later on. Furthermore, the error estimates that are induced by large degrees might dominate some behavior that occur in the large part of graphs with small degrees. Numerous problems remain unsolved several of which we mention here:

- In model A, we obtained the joint distribution of in- and out-degree. In general, it is true that $\lim_{t\to\infty}\frac{d_{i,j}^{joint}}{t}$ exists via the martingale theory. We denote it by $f(i,j)$. What is the asymptotic behavior of $f(i,j)$?

  Question: *Is there a simple asymptotic form of $f(i,j)$ (such as the in Theorem 1) for models B, C, D?*

- We only consider cases of adding nodes and edges at a time. This is consistent with some applications like a co-stars graph. However, for many other application, such as web graphs, there are often some destructions—deleting edges and nodes. Those destructions would affect the power law to some extent. Some cases were considered in Kleinberg et al.'s paper [22]. Simulations (and heuristic calculations) suggest that the in-degrees also follows the power law (see [22]). It would be of interest to analyze the more general model when deleting edges are allowed.

Question: *Can results for model C and D be extended allowing deleting nodes in the random process?*

References

[1] J. Abello, A. Buchsbaum, and J. Westbrook, *Proc. 6th European Symposium on Algorithms,* pp. 332–343, 1998.

[2] L. A. Adamic and B. A. Huberman, Growth dynamics of the World Wide Web, *Nature* , **401**, September 9, 1999, pp. 131.

[3] W. Aiello, F. Chung and L. Lu, A random graph model for massive graphs, *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, (2000) 171-180.

[4] Réka Albert, Hawoong Jeong, and Albert-Lászl´o Barabási, Diameter of the World Wide Web, *Nature* **401** (1999) 130-131.

[5] N. Alon and J. H. Spencer, *The Probabilistic Method*, Wiley and Sons, New York, 1992.

[6] Béla Bollobás, *Modern Graph Theory*, Springer-Verlag, New York, 1998.

[7] Albert-Lászl´o Barabási and Réka Albert, Emergence of scaling in random networks, Science **286** (1999) 509-512.

[8] A. Barabási, R. Albert, and H. Jeong, Scale-free characteristics of random networks: the topology of the world wide web, *Physica* A 272 (1999), 173-187.

[9] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, The degree sequence of a scale-free random graph process, *Random Structures and Algorithms*, **18**, (3), (2001), 279–290.

[10] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tompkins, and J. Wiener, “Graph Structure in the Web,” *proceedings of the WWW9 Conference*, May, 2000, Amsterdam. Paper version appeared in *Computer Networks* **33**, (1-6), (2000), 309-321.

[11] K. Calvert, M. Doar, and E. Zegura, Modeling Internet topology. *IEEE Communications Magazine*, **35(6)** (1997) 160-163.

[12] F. Chung and L. Lu, Connected components in random graphs with given degree sequences, preprint.

[13] F. Chung and L. Lu, The average degree in a random graph with a given degree sequence, preprint.

[14] C. Cooper and A. Frieze, A general model of web graphs. *http://www.math.cmu.edu/∼af1p/papers.html*, 2001.

[15] L. Egghe and R. Rousseau, *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*, Elsevier, 1990.

[16] P. Erdős and A. Rényi, On the evolution of random graphs, *Publ. Math. Inst. Hung. Acad. Sci.* **5** (1960), 17–61.

[17] P. Erdős and A. Rényi, On the strength of connectedness of random graphs, *Acta Math. Acad. Sci. Hungar.* **12** (1961), 261-267.

[18] A. A. Fairthore, Empirical hyperbolic distributions (Bradford Zipf Mandelbrot) for bibliometric description and prediction, *Journal of Documentation*, **25** (1969), 319-343.

[19] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the Internet topology, *Proceedings of the ACM SIGCOM Conference*, Cambridge, MA, 1999.

[20] N. Gilbert, A simulation of the structure of academic science, *Socialogical Research Online*, **2** (2), 1997.

[21] Zipf's law and prior distributions for the composition of a polulation, *J. Amer. Statis. Association*, **65** (1970), 1220-1232.

[22] J. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, The web as a graph: Measurements, models and methods, *Proceedings of the International Conference on Combinatorics and Computing*, 1999.

[23] M. Koenig and T. Harrell, Lotka's law, price's urn and electronic publishing, *Journal of the American Society for Information Science*, June 1995, 386-388.

[24] S. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, Trawling the web for emerging cyber communities, *Proceedings of the 8th World Wide Web Conference*, Toronto, 1999.

[25] S. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, Extracting large-scale knowledge bases from the web, *Proceedings of the 25th VLDB Conference*, Edinburgh, Scotland, 1999.

[26] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, Stochastic models for the Web graph, to appear in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* (FOCS 2000).

[27] A. J. Lotka, The frequency distribution of scientific productivity, *The Journal of the Washington Academy of the Sciences*, **16** (1926), 317.

[28] Linyuan Lu, The Diameter of Random Massive Graphs, *Proceedings of the Twelfth ACM-SIAM Symposium on Discrete Algorithms* (SODA 2001), 912-921.

[29] Tomasz Łuczak, Sparse random graphs with a given degree sequence, *Random Graphs*, vol 2 (Poznań, 1989), 165-182, Wiley, New York, 1992.

[30] H. A. Makse, S. Havlin and H. E. Stanley, Modelling urban growth patterns, *Nature*, **377** (1995), 608-612.

[31] C. Martindale and A. K. Konopka, Oligonucleotide frequencies in DNA follow a Yule distribution, *Computer & Chemistry*, **20** (1) (1996), 35-38.

[32] B. B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman and Company, 1977.

[33] G. A. Miller, E. B. Newman and E. A. Friedman, Length-frequency statistics for written English, *Information and Control*, **1** (1958), 370-389.

[34] Michael Molloy and Bruce Reed, A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, Vol. **6**, no. 2 and 3 (1995). 161-179.

[35] Michael Molloy and Bruce Reed, The size of the giant component of a random graph with a given degree sequence, *Combin. Probab. Comput.* **7**, no. (1998), 295-305.

[36] L. J. Murphy, Lotka's law in the humanities? *Journal of the American Society for Information Science*, Nov.-Dec. 1973, 461-462.

[37] P. Raghavan, personal communication.

[38] Newman, M., Strogatz, S., and D. Watts, "Random Graphs with Arbitrary Degree Distribution and Their Applications," submitted to Physical Review E.

[39] Center for Next Generation Internet, http://www.ngi.org/trends/TrendsPR0002.txt

[40] R. Rousseau and Q. Zhang, Zipf's data on the frequency of Chinese words revisited, *Scientometrics*, **24** (1992), 201-220.

[41] M. Schroeder, *Fractals, Caos, Power Laws*, W. H. Freeman and Company, (1991), 35-38.

[42] Z. K. Silagadze, Citations and the Zipf-Mandelbrot's law, *Complex Syst.* **11** (1997), 487-499.

[43] H. A. Simon, *Models of Man, Social and Rational*, New York, Wiley, 1957.

[44] , J. Tuldava, The frequency spectrum of text and vocabulary, *Journal of Quantitative Linguistics*, **3** (1996), 38-50.

[45] B. Waxman, Routing of multipoint connections. *IEEE Journal on Selected Areas in Communication* **6(9)** (1988) 1617-1622.

[46] N. C. Wormald, The asymptotic connectivity of labelled regular graphs, *J. Comb. Theory* (B) **31** (1981), 156-167.

[47] N. C. Wormald, Models of random regular graphs, *Surveys in Combinatorics*, 1999 (LMS Lecture Note Series 267, Eds J. D. Lamb and D. A. Preece), 239–298.

[48] G. U. Yule, *Statistical Study of Literary Vocabulary*, Cambridge Univ. Press, 1944.

[49] E. Zegura, K. Calvert, and M. Donahoo, A quantitative comparison of graph-based models for Internet topology. *IEEE/ACM Transactions on Networking*, **5** (6), (1997), 770-783.

[50] G. K. Zipf, *Human behaviour and the principle of least effort,* New York, Hafner, 1949.