Statistical inference on random graphs

by

Peter Hussami

Submitted to

Central European University

Department of Mathematics and its Applications

In partial fulfillment of the requirements for the degree of Doctor of Philosophy

CEU eTD Collection

Supervisor: Professor Gábor Tusnády

Budapest, Hungary

2010

Abstract

The main subject for this dissertation is statistical analysis of graphs. There are some general methodological musings, but the work focuses mostly on a specific problem: measuring prefential attachment in power-law degree distribution graphs and all the considerations that are involved. Namely, developing generating models and finding likelihood values for graphs. There is also heavy focus on uniform distribution over a class of graphs with a fixed degree sequence, which leads to the birth of a model where a graph's likelihood depends solely on its degree sequence.

CEU eTD Collection

Acknowledgements

I owe thanks to many people for their help. First and foremost among them is Gábor Tusnády. He is / was my supervisor and more: a great teacher and indeed a true torch-bearer.

The results of the thesis were discussed in many seminars and informal discussions with participants Réka Albert, Albert-László Barabási, Marianna Bolla, Villő Csiszár, Péter Erdős, Márton Ispány, János Komlós, György Michaletzky, István Miklós, Tamás Móri, Miklós Simonovits, Endre Szemerédi, and Tamás Vicsek, many thanks to them. I also thank Michael Krivelevich for giving the work a numerical push.

I thank the Central European University, the Rényi Institute and the Hungarian Academy of Sciences for providing me with the conditions for my work.

The thesis is to some extent a record of my development over the past five years – I owe credit to some people outside the field. My family was always beside me, come sun or rain. I thank my two consciences, Dávid and Thomas for being there. I cherish my times with the Work Club and Mid-week Sabbath Group for our work on liberating ourselves from the constraints of the greatest oppressor, language. I hope Kinga and Péter will one day get the helicopter they have been dreaming of. And I greet the shiny sunflower Hannah with a wink of an open eye.

CEU eTD Collection

Table of Contents

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

Chapter 1

Introduction

Our ancestors around the turn of the 20th century put tremendous effort into formulating mathematics as an axiomatic system of thought. The art has since then offered to anyone who wished to cultivate it the luxury of laboratory conditions, free of the imperfections of the surrounding world. Many of the problems mathematicians have worked on since that time have of course been inspired by "real-world" phenomena, but most results are formulated in a sterile environment, and anyone who wants to apply them rigorously must see to it that the necessary axioms are upheld.

Statistics is far less pure. True, there are beautiful statistical results in the literature, but it is by nature an applied branch of mathematics, and as soon as it is used for investigating a non-trivial problem, several questions arise. How to take an unbiased sample, for instance, can easily turn into a difficult problem, just as the question of what statistics to use for finding relevant patterns in the input data set. One needs only to think about, say, sociological studies to see how these seemingly minor questions influence the conclusions one might draw through "scientific" means, and might thus contribute to misunderstandings about the world we live in.

CEU eTD Collection

That arises most likely from the general perception that statistics is a toolkit for proving / disproving certain statements through numbers. In other words, its main function is pattern matching. That is of course correct if one considers statistical work to be only parameter estimation and hypotheses testing, as it is taught in most schools. But to the more motivated student, it is an art of pattern finding, rather than matching, and one can apply these methods in areas that go far beyond the realm of mathematics. That is very broad, but that is exactly the point: we learn to learn how to find some order in chaos, while relying as little on our (usually unreliable) perceptions as possible.

Some of the hidden patterns are easily guessed if there is some information about the data. For example, financial stock information should likely be affected by company sector and size. Behavioral patterns should relate to an individual's family background, their ability to speak foreign languages, and so on. But what other, non-trivial hidden patterns are there that might cluster the data? How should we go about finding them? How should we treat them? Several questions, without a universal answer. A particular set of tools may answer a particular mathematical problem, but in practice it is often very difficult to understand the meaning of patterns, when one gets to examine new data.

This study itself concerns investigating random graphs. Our ultimate aim would be to set up a graph lab where all possible random graphs in the world are classified according to some properties deemed generally important. But that is the work of a lifetime, and goes far beyond the scope of this dissertation. We will only list a number of properties that seem relevant for most graphs, and then proceed to a careful investigation of a relatively narrow area.

The amount of previous statistical work on graphs pales in comparison to graph theoretical work, but it does exist. Aldous has been a pioneer in the field (see for instance [1], or more on his web site), with recent results on noisy graphs by Bolla ([5]), on parameter estimation by Bickel and Chen ([4]), on the spectral properties of graphs generated by the

CEU eTD Collection

threshold model by Ide, Konno and Obata ([35]), and others. Tusnády is one of those who invested considerable effort into investigating random structures through statistical lens – on spectral properties (with Bolla, [6]), on general inference methodology (with Csiszár and Rejtő, [22]), on biological applications – e.g. on cortical networks (with Nepusz, Négyessy and Bazsó, [50]), or on matrix approximation applied specifically to mortality data (with Ispány, Michaletzky, Reiczigel, P. Tusnády and Varga, [36]).

The structure of the dissertation is as follows. In the next chapter we will introduce some earlier results on plain Erdős-Rényi and power-law graphs, then proceed to investigate methods that are useful for the statistical analysis of graphs. Then from chapter 4 onward we will focus on the following problem (indeed, our main problem): it has been established that if a graph is built up via some so-called "preferential attachment" mechanism, then it will have a power-law degree distribution. The reverse implication is generally taken for granted, but it is not so clear. In fact, it is so not clear that we will show in chapters 4 and 5 that it fails completely. In chapter 5 we will introduce some random graph models that produce power-law distributions, but this work will really be crowned by chapter 6, where we build a model that creates uniformly distributed graphs given a degree sequence.

1.1 A note on the notations

We intend to use consistent terminology and notation throughout the dissertation. For easier reading, we will label each layer of the text as follows:

- **[E]** - exact results
- **[H]** - heuristic results
- **[N]** - numerical results
- **[O]** - other authors' unpublished results

CEU eTD Collection

- **[S]** - survey of published results
- **[T]** - theoretical considerations

Here is a list of all frequently used entries here, with a short explanation attached:

- A graph G with n vertices and m edges will be notated as $G(n, m)$. A graph G with n vertices and some parameter relevant in the given context κ will be $G(n, \kappa)$.
- The vertex set of graph G will be denoted as $V(G)$, the edge set $E(G)$.
- The probability of an edge between vertices i and j of a graph will be p_{ij},
- The indicator variable for an edge between vertices i and j of a graph is ε_{ij},
- The degree of vertex i will be d_i or $deg(i)$. If this value cannot be assumed to be an integer, it might be notated as f_i,
- The likelihood of a graph: $L(G)$,
- $\mu(G)$ will stand for the class of graphs with a degree sequence identical to that of G,
- κ will be used as a scaling factor for the degrees of a graph's vertices. For most graphs $G(n, m)$, the relationship $m = \frac{1}{2}\kappa n$ will hold, but not everywhere.
- In the context of a random graph model, a nascent graph is simply a graph born out of that model.

CEU eTD Collection

Chapter 2

Background

[S] Random graph theory started in 1959 with a series of papers by P. Erdős and A. Rényi ([30], [29], [31]). They introduced two models with very interesting properties, but the graphs generated by these models were inadequate for describing some networks observed in real life. Networks became an area of intense research in the 1990s. Watts and Strogatz introduced a so-called "small world" model in 1998 ([57]), which aimed to modify the Erdős-Rényi models to better describe existing networks. In 1999, Albert and Barabási proposed a mechanism for building graphs that closely resembled real networks ([2]), notably in the power-law degree distribution. The publication was soon followed by another in similar vein, this time by Albert, Barabási and Jeong ([3]). The term "scale-free graph" was invented to describe graphs with such a distribution.

In this chapter we will introduce some of the major results in the field of random graph theory, with special attention to Erdős-Rényi ("ER") and Albert-Barabási ("AB") model graphs.

CEU eTD Collection

2.1 The Erdős-Rényi model(s)

There are two random graph models bearing the name "Erdős-Rényi":

$\mathcal{G}_{n,p}$: the set of graphs with n vertices, with each edge is present independent of the others with probability p,

$\mathcal{G}_{n,m}$: the set of graphs with n vertices, and m edges, the edge set picked uniformly over all $\binom{\binom{n}{2}}{m}$ combinations.

Maturally, $\mathcal{G}_{n,p}$ and $\mathcal{G}_{n,m}$ are graph classes and not graphs.

Edges in $\mathcal{G}_{n,p}$ are independent and equiprobable, which makes this model more calculation-friendly than $\mathcal{G}_{n,m}$. In $\mathcal{G}_{n,m}$, however, one can prescribe the number of edges the graph should have – increasing m gradually corresponds to the evolution of the graph.

Suppose $G \in \mathcal{G}_{n,p}$ and $p = \frac{m}{\binom{n}{2}}$. Then by the law of large numbers, $|E(G)| \approx m$ almost surely, and the two models exhibit similar behavior. There will be various occasions throughout this work where we will not care which model generated a graph, we will simply call it Erdős-Rényi.

2.1.1 Connectedness

Erdős and Rényi found that various properties of the model had sharp thresholds. Connectedness is perhaps the most fundamental one. Here we will mention two of their results.

First, they showed in [29] that for $G \in \mathcal{G}_{n,m}$, $m = \frac{1}{2}n \log n$ is a sharp threshold for connectedness.

Namely, for graph $G(n,m)$ with $m = \frac{1}{2}n \log n + cn$, $G(n,m)$ is connected with probability

$$\lim_{n\to\infty} P(G \text{ is connected}) = e^{-e^{-2c}}. \tag{2.1}$$

Depending on the value of c, the graph will be almost surely connected, or disconnected.

Second, in [30], a sharp threshold was given for the emergence of a giant connected

CEU eTD Collection

component in the graph. More precisely, for $n \to \infty$ and $p = \frac{c}{n}$, the following behavior is observed:

- For $c < 1$, the largest component has $O(\log n)$ vertices almost surely,
- for $c > 1$, there will emerge a giant component containing a positive fraction of the vertices, while all other components will have size $O(\log n)$.

These results inspired a lot of research. The article from Janson, Knuth, Łuczak, Pittel ([37]) is just one example, there are many more.

2.1.2 Degree distribution

In $\mathcal{G}_{n,p}$, the edges leaving a given node v are generated by $n-1$ flips of a p-probability coin, therefore:

$$P(d(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

For $n \to \infty$ and $np \to \lambda > 0$ constant, the degree distribution is Poisson:

$$P(d(v) = k) = \frac{(np)^k e^{-np}}{k!}.$$

2.1.3 Diameter

The diameter of a graph is defined as $diam(G) = max\{d(v,w)|v,w \in V(G)\}$, where $d(v,w)$ is the length of the shortest path between vertices v and w. If G is not connected, $diam(G) = \infty$ by definition.

The diameter is a monotone function on graphs, i.e. if G and H are graphs with $V(G) = V(H)$ and $E(G) \subset E(H)$, then $diam(G) \geq diam(H)$.

CEU eTD Collection

Bollobás shows in [7] that for a sufficiently dense graph $G \in \mathcal{G}_{n,p}$

$$diam(G) = \frac{\log n + \log\log n + \log 2 + O(1/\log n)}{\log pn}$$

almost surely. Specifically, for $c = \frac{n^2}{exp(p^d n^{d-1})}$ and $\frac{pn}{(\log n)^3} \to \infty$:

$$\lim_{n\to\infty} P(diam(G) = d) = e^{-\frac{c}{2}},$$

and

$$\lim_{n\to\infty} P(diam(G) = d+1) = 1 - e^{-\frac{c}{2}}.$$

2.1.4 Chromatic number

A proper vertex-coloring, or, for short, "coloring" of a graph G is a coloring on the graph's vertices such that any pair of adjacent vertices have a different color. The chromatic number of G is defined as the smallest number of colors required for coloring G, and is usually denoted as $\chi(G)$.

Regarding $\mathcal{G}_{n,p}$, we must distinguish between dense and sparse graphs. There is a stronger estimate available for dense graphs, but it fails for low edge densities.

First, the dense case. Suppose $G \in \mathcal{G}_{n,p}$, $p > n^{-\lambda}\ \forall\lambda > 0$ and let $b = \frac{1}{1-p}$. McDiarmid's 1989 result ([47]) shows that in this case

$$\chi(G) = \frac{n}{2\log_b n - 2\log_b\log_b n + O_C(1)}.$$

This result was an improvement upon earlier results of Shamir and Spencer ([53] and Bollobás ([8]).

In case p tends to zero too fast, the considerations that led the above result are not applicable. Using Frieze's argument ([32]), Łuczak was able to provide bounds in the sparse

CEU eTD Collection

case in his 1991 paper ([44]).

$\forall G \in \mathcal{G}_{n,p}, \exists C_0$ such that $\forall p = p(n)$ with p enclosed in the range $\frac{C_0}{n} \leq p \leq \log^{-7} n$, then asymptotically almost surely

$$\frac{np}{2\log np - 2\log\log np + 1} \leq \chi(G) \leq \frac{np}{2\log np - 40\log\log np}.$$

Note that for higher values of p, the previous estimate gives stronger bounds.

2.1.5 Cliques

A clique in graph G is defined as a complete subgraph of G. Specifically, let K_r denote a complete graph of order r. $k_r(G)$ will describe the number of distinct K_r-s in G. Let $\omega(G) = max\{r : \exists K_r \subset G\}$.

There are several reasons one might be interested in studying clique behavior of random graphs. For instance, if K_r is a maximum-sized clique in G, then the chromatic number, $\chi(\overline{G}) \geq r$, and K_r corresponds to an independent set in $\overline{G}$, the complement graph of G.

There has been a number of publications about the subject, among them Moon and Moser ([48]), Erdős ([27]), Grimmett and McDiarmid ([34], Matula ([45], [46]), and Bollobás-Erdős ([10]). Matula shows in [46] that as $|V(G)| \to \infty$, $\omega(G)$ takes one of at most two values with probability 1.

Let $b = \frac{1}{p}$, $d = 2\log_b n - 2\log_b\log_b n + 2\log_b(e/2) + 1 + o(1)$.

Then, $\forall G \in \mathcal{G}_{n,p}, 0 < p < 1, \varepsilon > 0$:

$$\lim_{n\to\infty} P(\lfloor d - \varepsilon \rfloor \leq \omega(G) \leq \lfloor d + \varepsilon \rfloor) = 1.$$

CEU eTD Collection

2.1.6 Subgraphs

The famous 1960 paper by Erdős and Rényi ([30]) discusses the subgraph containment problem at length. They provide threshold conditions for a random graph for containing certain small subgraphs. Namely, they show that:

If $k \geq 2$ and $k-1 \leq l \leq \binom{k}{2}$ are positive integers, and $\beta_{k,l}$ is an arbitrary non-empty class of connected graphs with k vertices and l edges, then $\forall G \in \mathcal{G}_{n,m}$ (or $G \in \mathcal{G}_{n,p}$ with $|E(G)| = m$):

$$P(\exists H \in \beta_{k,l} : H \subset G) = \begin{cases} 0, & \text{for } m = o(n^{2-\frac{k}{l}}) \\ 1, & otherwise \end{cases}$$

up to isomorphism over $\beta_{k,l}$.

Simple substitution yields that

- the threshold that G contains a tree of order k is $n^{\frac{k-2}{k-1}}$,
- the threshold that G contains a connected subgraph H with $|V(H)| = |E(H)| = k$ for arbitrary $k \geq 3$ is n,
- the threshold that G contains a cycle of order k for any $k \geq 3$ is n,
- the threshold that G contains a clique of order k for any $k \geq 3$ is $n^{\frac{2k-4}{k-1}}$.

2.2 The Albert-Barabási model

A quote from the 1960 Erdős-Rényi paper ([30]): 'It seems plausible that by considering the random growth of more complicated structures (e.g. structures consisting of different sorts of "points" and connections of different types) one could obtain fairly reasonable models of more complex real growth processes (e.g. the growth of a complex communication net

CEU eTD Collection

consisting of different types of connections, and even of organic structures of living matter, etc.).'

It seems the authors understood that the random graph models they worked with had their limitations as far as practical applications were concerned. Over the next decades, "real-life" networks were occasionally studied in detail, but it took until the 1990s that research sped up dramatically.

It became clear in many of these cases that the Erdős-Rényi model cannot provide an adequate description. The reason is instantly obvious: the degrees in the ER models have identical expected values, with a very narrow degree distribution. Many real-life networks, however, seem to carry a few degree-rich nodes and many degree-poor ones. So it seemed natural to investigate broader distributions, such as

$$P(deg(v) = k) \sim k^{-\lambda} \text{ for some } \lambda.$$

Albert, Barabási and others found that this so-called "power-law" distribution held well across a host of different real-life networks, for example:

- Collaboration graph for movie actors. Actors are represented by the nodes, and two nodes are connected by an edge if the corresponding actors have collaborated in a film. In this graph, $\lambda_{actor} = 2.3 \pm 0.1$.
- The World Wide Web. Each vertex represents a document, with a (directed) edge runs between them if one document has a link to the other. They found that $\lambda_{WWW} = 2.9 \pm 0.1$.
- The electrical power grid of the western United States. The vertices are generators, transformers, substations and the edges are the high-voltage lines between them. $\lambda_{power} \approx 4$ fit well.

CEU eTD Collection

- Scientific citations. The vertices are publications, the (directed) edges are references between them. It was found that $\lambda_{citation} = 3$.

- The number of sexual partners on a Swedish sample by Liljeros et al. ([43]). The study settled for $\lambda_{men} \approx 1.6$ and $\lambda_{women} \approx 2.1$. Their data suggests (again!) that men have more partners than women. And here we are, writing about graphs whose edges are supposed to have two end vertices...

All the above results are cited from Albert and Barabási ([2]) or a reference therein, unless otherwise noted.

Albert and Barabási pointed at two observations about such networks. First, that their vertex set generally grows over time, so any model that aims to describe them should consider the mechanism of how a new vertex is introduced to the system. And second, that a newly added vertex is more likely to connect to vertices that are already degree-rich. These two considerations motivated them to develop a model that leads naturally to a power-law distribution.

We will now describe the original Albert-Barabási model, and reproduce the authors' heuristic computations. In the next chapters we will provide a paraphrased version and analysis.

Consider starting graph $G_0(c, e)$. We operate in discrete time. For all $t \in \{1, 2..\}$, we add a new vertex v_{c+t} to G_{t-1}, which connects to exactly m ($\leq c$) of G_{t-1}'s nodes. So G_t has $c+t$ nodes and $e+mt$ edges. Denote by $\{w_1, \ldots, w_m\}$ the neighbors for v_{c+t+1} at time $c+t+1$. $\{w_1, \ldots, w_m\}$ is selected by the following method:

For picking w_k ($k = 1, \ldots, m$):

Suppose $X \in U(0, 1)$, and let $A_k = \{1, 2 \ldots, t+c\} \setminus \bigcup_{i=1}^{k-1} w_i$, an ordered set, with the elements indexed $A_{k,1}, \ldots, A_{k,t+c}$. Let $F(k) = \sum_{i \in A_k} d_i$, and $S(j) = \sum_{i=1}^{j} deg(A_{k,i})/F(k)$. Then w_k is the index of an element in A_k such that $S(w_k) < X$, but $S(w_{k+1}) \geq X$. For

CEU eTD Collection

$X = 1$, $w_k = |A_k|$. In brief, we are picking $w_1, \ldots, w_m$ from the distribution defined by $\{d_i\}$ without replacement.

Clearly,

$$P(N(v_i, t) = \{w_1, \ldots w_m\}) = \frac{d_{w_1}}{2(e+mt)} \frac{d_{w_2}}{2(e+mt) - d_{w_1}} \cdots \frac{d_{w_m}}{2(e+mt) - \sum_{i=1}^{m-1} d_{w_i}}.$$

The paper assumes $t \to \infty$, i.e. $\sum_{i=1}^{m} d_{w_i} \ll e + mt$, which means $\frac{mt}{e+mt} \approx 1$ and $P(N(v_i, t) = \{w_1, \ldots, w_m\}) \approx \prod_{i=1}^{m} \frac{d_{w_i}}{2mt}$.

So $P(w \in N(v_i, t)) = 1 - P(w \notin N(v_i, t)) = 1 - (\frac{2mt - d_w}{2mt})^m \approx m \frac{d_w}{2mt} = \frac{d_w}{2t}$.

Now, for a continuous-time approximation, consider the equation

$$\frac{\partial d_i}{\partial t} = \frac{d_i}{2t}.$$

This differential equation is solved by

$$d_i(t) = m\sqrt{\frac{t}{t_i}},$$

where t_i is the time vertex i appeared in the system, i.e. i. So

$$P(d_i(t) < k) = P(i > \frac{tm^2}{k^2}) = 1 - \frac{m^2}{k^2}.$$

To get the probability density, let us differentiate:

$$P(d_i(t) = k) = \frac{2m^2}{k^3}.$$

The above computations are very much heuristic in nature, and make some bold generalizations, such as approximating the neighbor selection process with independent draws, but the results are asymptotically correct.

CEU eTD Collection

A paper by Bollobás, Riordan, Spencer and Tusnády point out some formal problems with the above definition ([12]). Indeed, it is easy to see that the model has weak implications for microstructure. Take the case $m = 2$ for an example. If we were to connect v_{c+t} to the end vertices of a uniformly selected edge, each new vertex would add a triangle to the graph, which is obviously not the case in the model. Such considerations might inspire a mathematician to tighten the model.

In the next chapter we will redefine, and then modify the model such that it is easier to use for computations, while maintaining its most prominent properties.

For the rest of this section, we will note some of the known properties of scale-free graphs as defined by the above model. These are only a small sample of all that has been published about scale-free graphs. Newman, Barabási and Watts compiled a collection of important results about scale-free graphs ([51]), Dorogovtsev and Mendes wrote an early summary of this graph class ([23]), and many other results on the subject are available on the Internet.

2.2.1 Merits and dangers - politics

The popularity of the model inspired a lot of research, most of it outside mathematics. It is perhaps not impolite to say that an entire industry has been built up on scale-free networks. Numerous scale-free models have been introduced since Albert and Barabási's first paper on the subject, not the least by these same authors. These models are slight modifications to the original, to explain phenomena not contained in earlier versions, like how a late-comer might be able to attract neighbors at an exceptionally high rate, and so on.

The existence of such broadly applicable mathematical theory provides strong guidelines to researchers across many fields, and thus inspires work. This is welcome. True, much of what is understood by those who apply it is just a formal version of common sense.

CEU eTD Collection

One does not need highly developed theories to understand that the spread of signals in relatively centralized real-life networks depends largely on the central nodes, which means the Internet's backbone hubs need to be strongly protected, information / diseases spreads faster on them, etc. But having such models can be helpful in quantifying properties of such networks.

The process is not without dangers, however. The models tend to work with the assumption that scale-free distributions express something about a hidden law of nature. In the quest of understanding and describing the world, one is often tempted to turn to simplifications, and the existence of such universal assumptions might lead one to explain phenomena within the scale-free framework even where it is inappropriate. E.F. Keller ([39]) and others find this worrying. It is hard to argue against some of the criticism: forcing a scale-free uniform on too wide a range of networks should not become a world religion. But even if some think these models are weak or irrelevant, they do describe the growth of certain networks quite well. We will show that the model has well-defined connection properties, and since says something meaningful about certain real networks, its existence is justified.

Thus, it inspires the student to hammer out the model's weaknesses, rather than to discard it altogether.

Let us now look at some of the basic properties of scale-free graphs.

CEU eTD Collection

2.2.2 Connectedness

Denote the number of components of a graph G by $C(G)$. In the AB model, the addition of a new vertex does not increase the number of components, $C(G_0) \geq C(G_t)$, $\forall t \in \{1, 2, \dots\}$. Thus, if G_0 is connected, so will be G_t.

Suppose $C(G_0) = k$. For $m = 1$, $\forall t \in \{1, 2, \dots\}$, $C(G_t) = k$, since each new vertex connects to exactly one component. For $m > 1$: let $\{H_i\}$ denote the components with

$\sum_{v\in H_1} d_v \geq \cdots \geq \sum_{v\in H_k} d_v \geq 1$. Let $A_j(t)$ be the event that H_1 is connected to H_j at time t, and A_j the event that they are ever connected ($j = 1, \ldots, k$). These two components would be connected only if a vertex connected to them both, thus

$$P(A_j(t)) \geq \frac{1}{2mt}(\frac{mt}{k} \cdot \frac{1}{2mt})^{m-1} = \frac{k}{t}(2k)^{-m}.$$

Denote $c = \frac{2k^m}{k}$. Then

$$P(\overline{A}_j) \leq \prod_{t=t_0}^{\infty} \frac{ct-1}{ct} \leq \prod_{a=1}^{\infty}(\frac{ct-1+a}{ct+a})^{1/c} \to 0.$$

That holds for all $j = 2, \ldots, k$, thus if the graph is large enough, it will be connected.

Indeed, an AB graph is disconnected only if $m = 1$ and $C(G_0) > 1$.

2.2.3 Degree distribution

Albert and Barabási's estimate ($P(deg(v) = k) \sim \lambda^{-3}$. Dorogovtsev, Mendes and Samukhin show in [24] that for an AB graph with m edges added in each step:

$$P(d_v = k) = \frac{2m(m+1)}{k(k+1)(k+2)}.$$

Bollobás, Riordan, Spencer and Tusnády have a similar, but stronger result for $0 \geq k \geq n^{1/15}$ in [12].

The popular notation for the above is

$$P(deg(v) = k) \sim k^{-3},$$

for a randomly selected vertex.

CEU eTD Collection

2.2.4 Diameter

Bollobás and Riordan introduced a model in [11] that is quite similar to the AB model, with the exception that loops are allowed. They showed that if G is such a graph with n vertices and m edges, then for all $m \geq 2$ and $0 < \varepsilon$ real

$$(1-\varepsilon)\frac{\log n}{\log\log n} \leq diam(G) \leq (1+\varepsilon)\frac{\log n}{\log\log n}$$

almost surely.

The upper bound fails for $m = 1$. For $m = 1$, the average pairwise distance is $O(\log n)$, see [25], p. 109.

For a collection of further results, and indeed a lot of knowledge about random graphs, one is advised to turn to the random graphs book by Bollobás ([9]), Durrett ([25]) and/or Janson-Łuczak-Ruciński ([38]).

CEU eTD Collection

Chapter 3

Methods, considerations

[**T**] In this chapter we will give an overview of some of the considerations for evaluating graphs. We cannot afford to go into great detail about all of them – some of what is written here can simply provide the basis or motivation for future work.

We will use some of these methods in the later, more numerical part of the dissertation, here we wish to focus more on the philosophy. We believe the learning process is like chasing butterflies: butterflies are fast and elusive. Even if we manage to catch one, we might kill it, and what is the value of a dead butterfly? Still, we learn from the process, and that might be more important than finding answers. Over time, the answers will start coming...

The list we provide below is of course incomplete, but we hope the reader might find some of it not only helpful, but also inspiring for further work.

CEU eTD Collection

3.1 Models, likelihood, Neyman-Pearson

Statistics education tends to focus on two areas: parameter estimation and hypothesis testing. What is often not mentioned, because it is implicit in the former, is model estimation. Statistical data is always assumed to be generated by some model, and the parameters are basically parameters for that model. True, for simple models this distinction may not be necessary, but it would help students learn to be more critical with data.

So once we have settled for a model, we can look at the graph realization and try to derive its likelihood as per the model. This can be a very difficult task, if the model is complex (and graphs are reasonably complex), but there are great rewards if the attempt bears fruit. First, different realizations are comparable, and second, by the Neyman-Pearson lemma, different models are comparable. In practice, this means that we can make perturbations on the original data, and if the likelihood function changes significantly as a result of the perturbation, we have found something through a test of maximum power.

While this is not always achievable, it is a nice ideal to follow.

3.2 Degree distribution

The degree distribution is a property typical of every graph. There are numerous ways of comparing the empirical distribution and theoretical distributions. In practice these methods often aim at transforming the data in a way that a line should fit it well. For instance, the following approach is apparent in the original Albert-Barabási paper ([2]):

Suppose $P(d = k) \sim k^{-\lambda}$. Then

$$\log P(d = k) = -\lambda \log k + b,$$

so $\log P(d = k)$ should form a line with slope $-\lambda$ as a function of $\log k$.

CEU eTD Collection

In practice, many of the real-life graphs cited as scale-free show this linear relationship only in mid-range, and fail to fit near the head and tail of the distribution. More seriously, while the expected degree distribution associated with a random graph model might be known, the probability distribution of individual vertex degrees might not be identical, so as much as a nice picture might lead us to think we found a fitting, we should remain skeptical. So we are tempted to find some alternative methods for fitting distributions.

[**E**] Below is a small lemma for the discrete version of the probability integral transform.

Lemma 1. *Let D be a discrete distribution on m points, such that for $X \in D$, $P(X = i) = p_i$ ($i \in \{1, \ldots, m\}$), and let $r_i = \sum_{j=1}^{i} p_i$ ($i \in \{0, \ldots, m\}$). Let Y be a random variable conditioned on X, with*

$$P(Y \leq t|X) = \begin{cases} 0 & \text{if } t < r_{X-1} \\ 1 & \text{if } t \geq r_X \\ \frac{t - r_{X-1}}{r_X - r_{X-1}} & \text{otherwise} \end{cases}$$

Then $Y \in Unif(0, 1)$.

Proof. *Define $k(t)$ on $[0, 1]$ as the maximal integer such that $r_{k(t)} \leq t$. Then*

$$P(Y \leq t|X = i) = \begin{cases} 1 & \text{if } X < k(t) \\ \frac{t - r_{X-1}}{r_X - r_{X-1}} & \text{if } k(t) \leq X < k(t) + 1 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$P(Y \leq t) = \sum_{i=1}^{m} P(Y \leq t|X = i)P(X = i) =$$

$$(\sum_{i=1}^{k(t)-1} p_i) + p_{k(t)} \frac{t - r_{k(t)-1}}{r_{k(t)} - r_{k(t)-1}} = r_{k(t)-1} + p_{k(t)} \frac{t - r_{k(t)-1}}{p_{k(t)}} = t,$$

CEU eTD Collection

and the statement follows. □

[T] The above lemma allows us to fit the degree distribution very easily. Given a graph G with n degrees and m vertices, and a model M that is believed to have generated it, one needs to simple run the model enough times in order to attain an empirical degree distribution to each vertex. By this lemma, each of these empirical distributions can be converted to uniform on $(0, 1)$, so the fingerprint of the graph will be a sample of size n from $Unif(0, 1)$, if M is indeed the generating model, not otherwise. The uniform distribution can then be tested by some standard probes, say, the Kolmogorov test. This is of course not completely correct, since the degree numbers are not independent, but since their dependence is weak, this method gives a close approximation.

3.3 Uniform sampling from the degree distribution class

All the graph models we will look at have a degree distribution typical of them. This means that if we are given a graph $G(n, m)$ whose degree distribution is different from what the hypothesized model $\mathcal{M}$ would allow, it is grounds for us to reject the hypothesis that the $G(n, m) \subset \mathcal{M}$.

The other case is more interesting. Suppose $G(n, m)$'s degree distribution is in line model $\mathcal{M}$. Does this mean we accept the model?

We should not rush to judgment about this. True, the degree distribution test did not distinguish the graph from the model, but that was just one test. Suppose the $G(n, m)$ is in fact not from $\mathcal{M}$. Then there exists some statistic K that can distinguish it. If we are very lucky, $K(G(n, m))$ is very different from K on a graph born out of $\mathcal{M}$, and that is enough. This, however, is rare luck, we need a more reliable method.

CEU eTD Collection

Suppose $\mu(G(n,m))$ is the set of all graphs whose degree distribution is identical to that of G. If we can take a uniform sample H from $\mu(G)$ and find that $K(H)$ is significantly different from $K(G)$, we can declare victory. This method will always work, provided G stands out from $\mu(G)$ under K.

This opens the question of what uniform distribution on $\mu(G)$ is, i.e. what $\mu(G)$ exactly is. It makes a difference, for instance, if we consider vertex labels. Additionally, one might define a problem in a way that will introduce further considerations – e.g. we will define a labeled tree in the next chapter such that for every node there is ordering between the children. In this case, sampling uniformly could be quite a difficult task.

Another example: graphs from the Albert-Barabási process are by nature labeled, so much so that given a labeling, one could clearly tell whether a graph is Albert-Barabási or not. Yet, this labeling is not available when the graph is evaluated, so for all practical purposes, the graph must be considered unlabeled.

We will give this problem thought where we need it for solving a problem. The point we wish to make here is simply that a well-defined problem carries with it a definition of a uniform distribution, which can be a very helpful basis of comparison for solving the problem.

3.4 Labeling

Some graphs, like those from an AB model, evolve in time, and therefore have a natural labeling. These models may assign edge probabilities conditionally independently in every step, and thus reconstructing the original labeling would mean we can attain an accurate likelihood value for the graph. This motivates us to examine the problem of labeling.

The most natural idea, to consider all possible labelings and pick one according to some constraints, say, maximum likelihood, are computationally unrealistic. So it is clear that

CEU eTD Collection

for a general degree distribution, we are helpless about this problem.

So let us consider scale-free graphs, and say our graph $G(n, m)$ is from the AB model. The general tendency in these graphs is that the earlier a vertex is connected, the higher its degree. Móri shows in [49] that even as $n \to \infty$, the maximum degree node's identity will change only a finite number of times with probability 1. Naturally, this must hold true for the subgraph generated by vertices $\{i, \ldots, n\}$, regardless of the choice of i. This implies that for any fixed i and an arbitrary $j > i$, only $O(1)$ of these pairings will be such that $d_i < d_j$, out of $O(n)$ choices for j as $n \to \infty$. Thus, labeling the vertices in decreasing order by their degree is a negligible deviation from the original labeling.

We will use this as a general method: anytime we are to examine a scale-free graph whose nascent vertex order is not known, we will start with sorting the vertices by their degree. Note that this method is not applicable in general, but it does work for scale-free graphs.

3.5 Preferential attachment

[S] The term preferential attachment refers to a graph construction mechanism where every time a new vertex is joined to the graph, its neighbors are selected by some kind of preference function. Namely, if we wish to join vertex v to graph G to form G', we will have

$$\forall w \in V(G) : P((w, v) \in E(G')) \sim f(deg(w)).$$

This is a paraphrased form of the following urn process: let urn k (A_k) represent the degree of v_k. In step t of the process, some balls are picked from urns $1, \ldots, t-1$ with probabilities proportional to $f(A_1), \ldots, f(A_{t-1})$, and exactly as many balls are added to urn t.

CEU eTD Collection

Krapivsky, Redner and Leyvraz showed in [41] that for

- $f(k) = k^a$, with $a < 1$: $P(deg(v) = k) \approx ck^{-a} \exp(-ck^{1-a})$,
- $f(k) = k^a$, with $a > 1$: the resulting graph will have one vertex with degree of order t, others $O(1)$,
- $f(k) = a + k$, with $a > -1$: $P(deg(v) = k) \sim ck^{(-3+a)}$.

It is a task in itself to tell if the model that built a graph included preference considerations, so estimating f might be very difficult. But without a proper estimate for f, we might not be able to have an accurate likelihood estimate. For likelihood computations, we will assume the simplest of preferential attachment models, namely $f(k) = k$, and show that it is not too great a sacrifice. But for this reason we will be using the term "preferential attachment" loosely, pointing to a sequential model's property that higher degree vertices have greater strength in attracting neighbors, and not require f to be linear.

3.6 Clustering

[T] Clustering on a graph is, in the broadest sense, a partition on the vertex set. The term "clustering" is typically used for a partitioning where vertex groups are formed from a set of vertices that are in some sense similar. It is important to note that this is only one of the possible ways to approach the subject of clustering.

One can also define partitioning in terms of cross-cluster behavior. The criterion for a vertex to belong to a cluster could be its distance to vertices in other clusters (according to some distance metric), or by some other regularity property.

To illustrate the difference, consider a graph with some distance metric defined over it, say, a real-world map. Clustering by geographic location is of the first type, and by the vertex degree is of the second type.

CEU eTD Collection

Let us briefly discuss these two approaches.

3.6.1 Clusters of "similar" vertices

This approach has very rich literature, so there is no need to go into technical details. The only consideration we can add is how to define a distance metric that is relevant for our graph. Hamming distance is one metric that comes to mind automatically, another one is the one implied by the process of drawing up the graph, where a metric space is set up automatically. The list could go on.

If we do not want to make a conscious decision on the metric space, we might want to set up some constrained process, and let the vertices find their own place, hoping there is a stationary state. For instance, we could randomly distribute the vertices along a spherical surface, and view edges as gravitational forces that pull the end vertices together. The surface is automatically metric, so once the stationary state is reached, clustering becomes a trivial problem.

There is no immediate summary of what this might do to a particular graph, it depends on the surface and the dynamic process, but once we have applied the method to a wide range of graphs, some of their properties might just stand out. And over time, one will develop an intuition for the features that seem relevant.

CEU eTD Collection

3.6.2 Clusters by cross-cluster behavior

One automatic idea that is inspired by the gravity example is to use "anti-gravity" instead of gravity between the graph's vertices. But the main point we would like to make in this section has to do with a very famous and deep previous result.

There has been a lot of interest in Szemerédi's regularity lemma in recent years (see references [54], [55]), with a host of analogous results across various fields of mathematics.

In the simplest terms, the lemma essentially states that for a given positive integer k, every graph with a large enough number of vertices (depending on k) can be divided into roughly identical-sized subsets such that the pairwise edge densities between the subsets is constant, save for some small error.

Here the regularity appears between the clusters, not within them. That means the graph could be relabeled such that its adjacency matrix would contain approximately identically dense blocks everywhere except in blocks along the main diagonal.

Szemerédi's result applies to dense graphs only. No analogous result of such strength has been found for sparse graphs, but some advances have been made, for example by Y. Kohayakawa ([40]).

3.6.3 Clustering, rank and eigenvalues

We will not spend much time with spectral theory, but let us remark here a small observation about the connection between rank, clusters and eigenstructures.

Let n and c be positive integers with $c \ll n$. Let P be the c x c matrix arising from the principal axis transformation:

$$\forall 1 \leq i \leq j \leq c : P_{ij} = \sum_{s=1}^{c} \lambda_s v_i(s) v_j(s),$$

for some scalars $\lambda_1, \ldots, \lambda_c$ and c-dimensional vectors $v_1, \ldots, v_c$.

Then construct function $\alpha(u)$ such that $\forall u = 1, \ldots, n$, $1 \leq \alpha(u) \leq c$. Let H be an n x n matrix such that

$$H_{u,v} = P_{\alpha(u),\alpha(v)} = \sum_{s=1}^{c} \lambda_s v_{\alpha(u)}(s) v_{\alpha(v)}(s).$$

Then $rank(H) \leq rank(P)$.

If $P_{\alpha(u),\alpha(v)}$ is relatively close to 0 or 1, then H's rows and columns can be reshuffled to

CEU eTD Collection

form uniform density blocks.

There is a number of interesting articles on the subjects, like the one by Bolla and Tusnády ([6]), Coja-Oghlan ([20]), Chung and Lu ([17], [16]), and Chung, Lu and Vu ([19], [18]).

3.7 Edge correlations

We might consider the process of edge formation over as follows:

Let $G(n, m)$ be a random graph, with $\{d_i\}$ the expected degree distribution of G ($i = 1, \ldots, n$). Let $\{p_i\}$ be a distribution over $\{1, \ldots, n\}$:

$$p_i = \frac{d_i}{\sum_{j=1}^{n} d_j}.$$

Let $\{E_j\}$ be a series of random variables representing the edges ($j = 1, \ldots, m$), with $E_j = (X_j, Y_j)$, with X_j and Y_j picked i.i.d. from the above distribution.

Then we can calculate the edge correlation as

$$corr(E_X, E_Y) = \frac{E(E_X E_Y) - E(E_X)E(E_Y)}{\sqrt{var(E_X)var(E_Y)}},$$

since $E(E_X) = E(E_Y)$.

3.8 Potentials

One of the problems that we have with graphs is that there is no trivial metric for them. We define the word "potential" in a very broad sense. Let us define it as a monotone measure of the graph, i.e. for potential P:

CEU eTD Collection

$$G \subset G' \Rightarrow P(G) \leq P(G').$$

$|E(G)|$ is perhaps the most simple potential, but the number of paths of a given length would also qualify as one, and so on, the possibilities are endless. Different potentials reveal different features of the graph, they are often invented along the way of solving a particular problem, and thus do not require our further attention.

Except for one, the S metric.

3.8.1 The S metric

[S] Li et al proposed a potential in their 2005 paper ([42]) they claimed would be very relevant for the class of scale-free graphs. This so-called "S metric" is defined as follows:

$$S(G) = \sum_{i,j \in E(G)} d_i d_j$$

The authors go on suggesting that this metric is be helpful in defining scale-free graphs. Namely:

Let G be a graph with a power-law degree distribution and scaling index λ. Let $\mu(G)$ be the class of graphs with a degree sequence identical to that of G. Then define $S_{max}(\mu(G))$ as:

$$S_{max}(\mu(G)) = max\{S(H) \mid H \in \mu(G)\}$$

.

And define:

$$T(G) = \frac{S(G)}{S_{max}(\mu(G))}$$

.

CEU eTD Collection

The authors claim that the value of $T(G) \approx 1$, at least if $\lambda \geq 2$, and propose that scale-free graphs should be defined as graphs where this T value is close to 1. We will show in section 5.4.4 that this is not true. Nonetheless, it is still a potential easy to comprehend and work with, and we will make heavy use of it.

3.9 Microstructures

"The devil is in the details" - the saying goes. One approach to understanding a graph is through its small subgraphs, which are not always diabolic. These are easy to grasp both analytically and computationally. Borgs, Chayes, Lovász, T. Sós and Vesztergombi have two elegant papers about the subject ([14], [13]), where they introduce a notion of convergent graph sequences based on small subgraphs.

If we are able to compute some asymptotic subgraph properties of a hypothesized graph model, we might use this as part of a test on the graph we are examining. This so-called "containment problem" is often an easy task, since it involves small subgraphs. Section 2.1.6 has a demonstration of some basic results for Erdős-Rényi graphs.

3.10 Distances, metrics

Given two general graphs, we might want to know how "similar" they are. This becomes a very difficult problem if these graphs are not labeled. There are some fast algorithms finding isomorphism between trees, and that means one can come up with a tree-to-tree distance metric with some effort, but the same does not hold for general graphs.

The graph isomorphism problem is known to belong to NP, but it is currently unknown whether it is NP-complete or not. The subgraph isomorphism problem is in fact NP-complete ([56]). That means coming up with a practical graph distance algorithm is by no

CEU eTD Collection

means a trivial matter.

Non-trivial does not mean unsolvable: Frieze and Kannan introduced a cut or rectangle distance in [33]. This distance satisfies all conditions for being a metric. The graph convergence idea developed by the Borgs, Chayes, Lovász, T. Sós and Vesztergombi articles mentioned in the previous section lay the groundwork for metricization. These ideas are very complex computationally, but the idea of a metric appears there. That implies a small rank graph that can be partitioned into a low number of clusters can be easily metricized in practice.

We have now introduced some general methods for examining graphs. Let us proceed to our main problem.

CEU eTD Collection

Chapter 4

Scale-free distributions: a tree

[E] Motivating problem: it is understood that if a graph is built through linear preferential attachment, it will have a power-law degree distribution. We would like to investigate the reverse implication. Formally, for graph G:

$$G \text{ was built through pref. att. } \Rightarrow \forall v \in G, P(deg(v) = k) \sim k^{-\lambda}, \tag{4.1}$$

for some λ. How about $\Leftarrow$?

To this end, we will set up a simple scale-free model that will allow us to compute this and other properties. Let us define a random process that yields a scale-free tree.

4.1 The model

Let G_1 be a simple labeled graph consisting of two vertices (0 and 1) and one edge. For all $k = \{1, 2, \dots\}$, graph G_{k+1} is obtained from G_k via the addition of a new vertex $k+1$, connecting it to one of the end vertices of an edge selected uniformly from $E(G_k)$. The end vertex v is chosen with probability $\frac{1}{2}$ each. Clearly, v is picked with probability

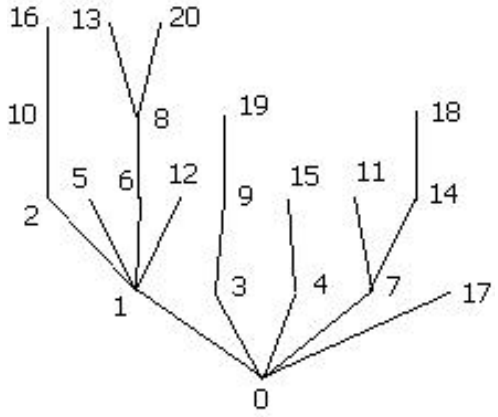
CEU eTD Collection

proportional to its degree in G_k, so this model is equivalent to an AB model with $m = 1$ and G_0 consisting of just an edge. See figure 4.1 for an illustration for such a tree.

Given a tree G, one might want to know how likely the above process could have created it. To paraphrase, we wish to understand the likelihood of each graph $G(n+1)$ as above.

First, let us examine the case where all vertices are strictly labeled.

Figure 4.1: An example tree, with $L = \frac{2!2!4!4!}{40!!}$



Proposition 1. *Let $G(n+1)$ be as above, and $\Lambda(G(n+1))$ be the likelihood of $G(n+1)$ with a fixed labeling. Then*

$$\Lambda(G(n+1)) = \frac{1}{(2(n-1))!!} \prod_{v \in V(G)} (d_v - 1)!. \tag{4.2}$$

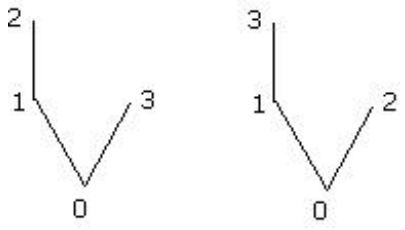
Proof. *Observe that when v_{k+1} is added to graph G_k, the choice of its neighbor is independent of all past and future neighbor choices. The probability of choosing $w \in G_k$ as the neighbor of v_{k+1} is exactly $\frac{d_w(k)}{2k}$. Repeating over $k = \{1, 2 \dots n-1\}$, the above result is obtained for $G(n+1)$.* □

The point of this exercise is to reconstruct the creation of an AB tree, which is not explicitly labeled, but contains some implicit labeling restrictions. Notably, given a root,

CEU eTD Collection

for any edge it is clear which of its end vertices is the parent and the child in the tree. Moreover, the order between any two children of the same vertex matters, since an "elder" child - the one with the lower label - is likely to have more offsprings than its younger siblings.

Figure 4.2: The smallest tree with multiple labelings



The likelihood of a tree G with a particular fixed labeling is as above. But that is only half a solution. As we can see in figure 4.2, a tree in general can have multiple labelings. Given G, we would not like to find out how many trees there are that are isomorphic to G, have the same root and observe G's order of siblings with respect to the root.

Definition. *Let G be a tree with labeling l, and U a subtree of G, with $root(U) = u$. We will call l conservative on G if $\forall U \subset G$, $\forall v \in U \backslash \{u\}$: $l(u) < l(v)$.*

Definition. *Let G be a tree with conservative labelings l and m, and U a subtree of G with $root(U) = u$. Let $v_1 \dots v_k$ be u's children. We will call m l-preserving if and only if $\forall U \subset G$, $\forall i, j \in \{1, \dots k\}, i \neq j$:*

$$l(v_i) < l(v_j) \Rightarrow m(v_i) < m(v_j).$$

In other words, m preserves the older-younger sibling relationships present in l.

Definition. *The offspring tree of vertex v in tree G is the subtree $G' \subset G$ such that $root(G') = v$.*

CEU eTD Collection

Definition. *The family F of a vertex v in labeled tree G with labeling l is the offspring tree of v and all its younger siblings, i.e.*

$$F(V) = \{w \in G \mid w \text{ a sibling of } v \text{ such that } l(w) > l(v)\}.$$

Theorem 1. *Let $G(n+1)$ be as above, a tree with $n+1$ vertices and n edges, and labeling l. Let $C(G,l)$ be the set of l-preserving labelings on $G(n+1)$. Denote by f_i the vertex family size of vertex i. Then*

$$|C(G,l)| = \frac{n!}{(\prod_{i=1}^{n} f_i)}. \tag{4.3}$$

Proof. *Consider an arbitrary vertex v_0 and its immediate children $\{v_1, \dots v_k\}$, indexed such that $l(v_1) < l(v_2) < \cdots < l(v_k)$. Denote by $T_0 \dots T_k$ the offspring trees of vertices $v_0 \dots v_k$, respectively. Let $M_i = |T_i|, i = \{0, \dots k\}$, and f_i the vertex family size of $v_1 \dots v_k$.*

In the case $k = 1$, we get the recursion $|C(T_0,l)| = |C(T_1,l)|$.

For $k \geq 2$: for any pair $i, j \in \{1, \dots k\}, i \neq j$, the number of l-preserving labelings on T_i and T_j is an independent factor of $|C(T_0,l)|$, which yields the recursion

$$|C(T_0,l)| = A(T_0) \prod_{i=1}^{k} |C(T_i,l)|,$$

with $A(T_0)$ to be determined.

l is conservative, which means for all trees $T_1, \dots, T_k$, its root must have the smallest label on that subtree. Moreover, $l(root(T_1)) < \cdots < l(root(T_k))$, which means:

$$A(T_0) = \binom{M_0 - 2}{M_1 - 1}\binom{M_0 - M_1 - 2}{M_2 - 1} \cdots \binom{M_k - 1}{M_k - 1} =$$

$$\frac{(M_0-2)!}{(M_1-1)!(M_0-1-M_1)!} \cdots \frac{(M_0 - 2 - \sum_{i=1}^{k-2} M_i)!}{(M_{k-1}-1)!(M_0 - 1 - \sum_{i=1}^{k-1} M_i)!} \frac{(M_k-1)!}{(M_k-1)!}$$

CEU eTD Collection

Observe that for all $a = 1, \ldots, k$*:* $M_0 - 1 - \sum_{i=1}^{a-1} M_i = f_i$*. After simplification we get*

$$A(T_0) = \frac{(M_0 - 2)!}{\prod_{i=1}^{k}(M_i - 1)! \prod_{i=2}^{k} f_i} = \frac{(M_0 - 1)!}{\prod_{i=1}^{k}(M_i - 1)! \prod_{i=1}^{k} f_i}.$$

For $k = 1$*, the above formula for* $A(T_0)$ *yields* $\frac{1}{f_1}$*, and if* T_1 *is a child,* $A(T_0) = 1$*.*

As the recursion unfolds, the factor $\frac{1}{(M_i-1)!}$ *is cancelled out by* $(M_i-1)!$ *in the numerator for vertex* i*'s leftmost child (*$i = 1, \ldots, k$*). This leads to*

$$|C(T_0, l)| = \frac{(M_0 - 1)!}{\prod_{v \in T_0 \setminus v_0} f_v}.$$

Substituting $v_0 = root(G)$*, we get*

$$|C(G, l)| = \frac{n!}{\prod_{i=1}^{n} f_i},$$

confirming the statement.

□

Note that the inclusion of the root node would yield

$$|C(G, l)| = \frac{(n+1)!}{\prod_{i=0}^{n} f_i},$$

which is identical, as $f_0 = n + 1$.

In summary, the likelihood that an $n + 1$ vertex tree T is isomorphic to T and has proper labeling with respect to T's root is

$$L(T) = \frac{n!}{(2(n-1))!! \prod_{i=1}^{n} f_i} \prod_{i=1}^{n} (d_i - 1)!. \tag{4.4}$$

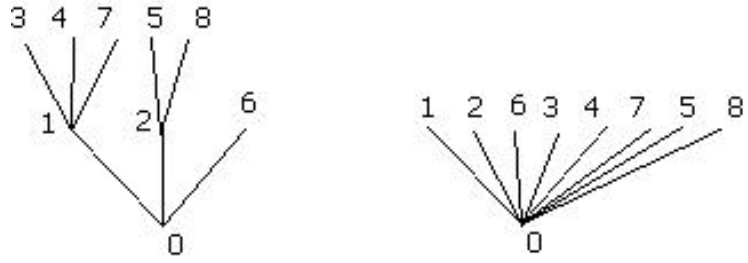
Proposition 2. *Let* $G(n+1)$ *be as above with labeling* l*,* $\Lambda(G)$ *be the likelihood of* $G(n+1)$

CEU eTD Collection

with a fixed labeling and $C(G) = C(G, l)$ the set of l-preserving labelings. Then $L(G) = \Lambda(G)C(G)$ is maximal if and only if all vertices are connected to one of the initial two vertices.

Proof. *The method will be to take an arbitrary subtree of depth 2 and show that L will increase if the tree is flattened to depth 1 (see figure 4.3).*

Figure 4.3: Flattening. The old labeling, while wrong, is kept for illustration



Let T be a subtree of G with depth 2. It is clear from equations (4.2) *and* (4.3) *that the only arithmetic effect any rearrangement of T will have on the values $\Lambda(G)$ and $C(G, l)$ are contained within T, as long as no vertices are added or deleted. For shorthand, we will use the notation $C(G) = C(G, l)$.*

Suppose the root of T is v_0, its k children are $v_1 \dots v_k$. Denote by M_i the size of v_i's offspring tree ($i \in \{1 \dots k\}$).

Case 1: $T = G$.

$$C(T) = \frac{(\sum_{i=1}^{k} M_i)!}{\prod_{i=1}^{k}(M_i - 1)! \prod_{i=1}^{k} \sum_{j=i}^{k} M_j},$$

and

$$\Lambda(T) = \frac{(k-1)! \prod_{i=1}^{k}(M_i - 1)!}{(2(\sum_{i=1}^{k} M_i) - 1)!!}.$$

Let T' be the "flat" version of T, i.e. a root and $\sum_{i=1}^{k} M_i$ children.

CEU eTD Collection

$$C(T') = \frac{(\sum_{i=1}^{k} M_i)!}{(\sum_{i=1}^{k} M_i)!} = 1,$$

and

$$\Lambda(T') = \frac{((\sum_{i=1}^{k} M_i) - 1)!}{(2(\sum_{i=1}^{k} M_i) - 1)!!}.$$

Thus,

$$\frac{\Lambda(T)C(T)}{\Lambda(T')C(T')} = \frac{(k-1)! \sum_{i=1}^{k} M_i}{\prod_{i=2}^{k} \sum_{j=i}^{k} M_j} \leq 1,$$

with equality holding only if all of $M_2, \ldots, M_k$ *is* 1. *The case when* $M_1 \geq 1$ *arbitrary, and all other* M_i*-s are* 1 *has all leaves of the tree dangling on the end vertices of the initial edge, as stated.*

Case 2: $T \neq G$.

Here we will compute $R_\Lambda(T)$ *and* $R_C(T)$, T*'s multiplicative contribution to* $\Lambda(G)$ *and* $C(G, l)$, *respectively. The formulas for* $\Lambda(G)$ *and* $C(G, l)$ *contain* n *in the denominator or numerator, here we will ignore them.*

$$R_C(T) = \frac{1}{\prod_{i=1}^{k}(M_i - 1)! \prod_{i=1}^{k} \sum_{j=i}^{k} M_j},$$

and

$$R_\Lambda(T) = \frac{k! \prod_{i=1}^{k}(M_i - 1)!}{(2(n-1))!!}.$$

Again, let T' *be the flattened version of* T.

$$R_C(T') = \frac{n!}{(\sum_{i=1}^{k} M_i)!} = 1,$$

CEU eTD Collection

and

$$R_\Lambda(T') = \frac{(\sum_{i=1}^{k} M_i)!}{(2(n-1)!!}.$$

Thus,

$$\frac{R_C(T)R_\Lambda(T)}{R_C(T')R_\Lambda(T')} = \frac{k!}{\prod_{i=1}^{k}\sum_{j=i}^{k} M_j} < 1.$$

This is a strict inequality, as T has depth two, and thus $M_j > 1$ for some j. Therefore, flattening the tree from near the leaves down to the root will always increase ΛC, all the way until all edges vertices $2, \ldots, n+1$ are connected to the first edge. □

It does not matter for the maximum of $\Lambda(G)C(G)$ how the edges are distributed between the two end vertices of the first edge. Suppose k of them are neighbors of v_0, and $n-k-1$ are neighbors of v_1. Then

$$\Lambda(G)C(G) = \frac{k!(n-k-1)!}{(2(n-1))!!} \cdot \frac{n!}{(n-k-1)!nk!} = \frac{(n-1)!}{(2(n-1))!!},$$

independent of k.

This implies that the maximum likelihood tree will not be a typical outcome for this scale-free tree model.

CEU eTD Collection

4.2 The probability distribution of the degrees

[O] We will use the urn model to compute the distribution of the tree's vertex degrees. Consider the following urn game: we have two urns, the initial weights of the first and second are p and r, respectively. In each round we pick an urn with a probability proportional to its current weight, and increase its weight by one. Regardless of the choice we make, we also increase the second urn's weight by one. Let X_n denote the first urn's weight after n

steps.

Thanks to Tamás F. Móri for the result below.

Lemma 2. *$\frac{X_n}{\sqrt{2n}} \to \zeta$ with probability 1, where all the following hold for $F(\zeta)$ ζ's distribution:*

- *(a) $F(\zeta) \sim \sqrt{UV}$, where U and V are independent and $U \sim \Gamma(\frac{p+1}{2}, \frac{1}{2})$, $V \sim \beta(\frac{p}{2}, \frac{r}{2})$*
- *(b) $F(\zeta) \sim V\sqrt{U}$, where U and V are independent and $U \sim \Gamma(\frac{p+r+1}{2}, \frac{1}{2})$, $V \sim \beta(p, r)$*
- *(c) For $r = 1$, $F(\zeta) \sim \sqrt{U}$, where $U \sim \Gamma(\frac{p}{2}, \frac{1}{2})$*
- *(d) For $r > 1$, $F(\zeta) \sim \sqrt{UV}$, where U and V are independent and $U \sim \Gamma(\frac{p}{2}, \frac{1}{2})$, $V \sim \beta(\frac{p+1}{2}, \frac{r-1}{2})$.*

Proof. *We will compute the moments of the limit distribution through martingales.*

Let k be a particular positive integer, and let

$$c_n = \prod_{i=0}^{n-1}(1 + \frac{k}{p+r+2i}) = \frac{\Gamma(\frac{p+r+k}{2}+n)\Gamma(\frac{p+r}{2})}{\Gamma(\frac{p+r+k}{2})\Gamma(\frac{p+r}{2}+n)} \sim \frac{\Gamma(\frac{p+r}{2})}{\Gamma(\frac{p+r+k}{2})} n^{\frac{k}{2}}.$$

Then $Z_n = \frac{1}{c_n}\binom{X_n+k-1}{k}$ is a martingale, since

$$E(Z_{n+1}|\mathcal{F}_n) = \frac{1}{c_{n+1}}\left[\binom{X_n+k-1}{k} + \frac{X_n}{p+r+2n}\binom{X_n+k-1}{k-1}\right] =$$
$$\frac{1}{c_{n+1}}(1 + \frac{k}{p+r+2n})\binom{X_n+k-1}{k} = Z_n,$$

where $\mathcal{F}_n$ is the appropriate generated σ-algebra.

Then

$$E(Z_n) = E(Z_0) = \binom{p+k-1}{k} = \frac{\Gamma(p+k)}{\Gamma(p)k!},$$

and

$$\lim_{n\to\infty} Z_n = \frac{2^{\frac{k}{2}}\Gamma(\frac{p+r+k}{2})}{k!\Gamma(\frac{p+r}{2})}\zeta^k.$$

CEU eTD Collection

$E(\zeta^k)$ bounded for all values of k, and thus the expected value of the above martingale's limit is the same as that of the martingale itself, which yields

$$E(\zeta^k) = \frac{\Gamma(\frac{p+r}{2})\Gamma(p+r)}{2^{\frac{k}{2}}\Gamma(\frac{p+r+k}{2})\Gamma(p)}.$$

Replacing k with $2k$ and using the duplication relation

$$\Gamma(2x) = (2\pi)^{-\frac{1}{2}}2^{2x-\frac{1}{2}}\Gamma(x)\Gamma(x+\frac{1}{2})$$

we get

$$E(\zeta^{2k}) = \frac{\Gamma(\frac{p+1}{2}+k)}{2^{-k}\Gamma(\frac{a+1}{2})} \cdot \frac{\Gamma(\frac{p+r}{2})\Gamma(\frac{p}{2}+k)}{\Gamma(\frac{p}{2})\Gamma(\frac{p+r}{2}+k)}.$$

The s-th moment of the distribution $\Gamma(a,\lambda)$ is $\frac{\Gamma(a+s)}{\lambda^s\Gamma(a)}$. The same for the distribution $\beta(a,b)$ is $\frac{\Gamma(a+b)\Gamma(a+s)}{\Gamma(a)\Gamma(a+b+s)}$, thus the above quantity decomposes as the product of the k-th moments of a $\Gamma(\frac{p+1}{2},\frac{1}{2})$ and a $\beta(\frac{p}{2},\frac{r}{2})$, which proves (a).

It also follows from the duplication relation that

$$\frac{\Gamma(\frac{p+r}{2})}{\Gamma(\frac{p+r+k}{2})} = \frac{\Gamma(\frac{p+r+k+1}{2})\Gamma(p+r)2^k}{\Gamma(\frac{p+r+1}{2})\Gamma(p+r+k)},$$

so

$$E(\zeta^k) = \frac{\Gamma(\frac{p+r+k+1}{2})}{2^{-\frac{k}{2}}\Gamma(\frac{p+r+1}{2})} \cdot \frac{\Gamma(p+r)\Gamma(p+k)}{\Gamma(p)\Gamma(p+r+k)}.$$

This is a product of the $\frac{k}{2}$-th moment of the $\Gamma(\frac{p+r+1}{2},\frac{1}{2})$ and the k-th moment of the $\beta(p,r)$ distributions, proving (b).

For $r=1$:

$$E(\zeta^{2k}) = \frac{\Gamma(\frac{p}{2}+k)}{2^{-k}\Gamma(\frac{p}{2})},$$

which is exactly the k-th moment of the $\Gamma(\frac{p}{2},\frac{1}{2})$ distribution, proving (c).

CEU eTD Collection

Finally, if $r > 1$, we can attain the product in a slightly different form:

$$E(\zeta^{2k}) = \frac{\Gamma(\frac{p}{2}+k)}{2^{-k}\Gamma(\frac{p}{2})} \cdot \frac{\Gamma(\frac{p+1}{2}+k)\Gamma(\frac{p+r}{2})}{\Gamma(\frac{p+1}{2})\Gamma(\frac{p+r}{2}+k)}.$$

Here the first factor is the k-th moment of the $\Gamma(\frac{p}{2}, \frac{1}{2})$ distribution, and the second is the k-th moment of the $\beta(\frac{p+1}{2}, \frac{r-1}{2})$ distribution, yielding (d). □

The distribution of the k-th vertex in our scale free tree is described by the case $p = 1$, $r = 2k - 1$. Lemma statements (a), (b) and (d) are applicable.

In the case $r = 1$, the probability density function is

$$f_{p,1}(x) = \frac{2x^{p-1}e^{-\frac{x^2}{2}}}{2^{\frac{p}{2}}\Gamma(\frac{p}{2})} = \frac{2^{\frac{p+1}{2}}\Gamma(\frac{p+1}{2})}{\Gamma(p)}.$$

Specifically, for $p = 1$, it reduces to

$$f_{1,1}(x) = 2\varphi(x).$$

And thus the distribution of d_1 is the standard normal absolute value distribution.

Lemma 3. *Denote by $\zeta_{p,r}$ the limit distribution with parameters p and r. Then:*

- *(a) For $p \to \infty$ and $r = o(\sqrt{p})$, $\sqrt{2}(\zeta_{p,r} - \sqrt{p}) \to N(0,1)$ in distribution.*
- *(b) For $p = 1$, $r \to \infty$, $\sqrt{r}\zeta_{p,r}$ converges in distribution to the exponential distribution with parameter 1 (this will be applicable to our tree model).*

Proof. *We will use statement (a) of the previous lemma for both statements.*

(a) So we have

$$\sqrt{2}(\sqrt{UV} - \sqrt{p}) = \sqrt{2V}(\sqrt{U} - \sqrt{p}) - \sqrt{2p}(1 - \sqrt{V}).$$

CEU eTD Collection

$\sqrt{2p}(1-\sqrt{V}) < \sqrt{2p}(1-V)$. *The expected value of which is* $\frac{p\sqrt{2p}}{p+r} \to 0$, *so the* $\sqrt{2p}(1-\sqrt{V}) \to 0$. *Therefore, the equation implies* $\sqrt{V} \to 1$.

Furthermore, $E(U) = p+1$, $D^2(U) = 2(p+1)$, *thus* $p^{-1}U \to 1$. *Then, by the central limit theorem:*

$$\sqrt{2V}(\sqrt{U}-\sqrt{p}) = \frac{2\sqrt{V}}{\sqrt{p^{-1}U+1}}\frac{U-p}{\sqrt{2p}} \to N(0,1).$$

(b) The distribution of U *is* $\Gamma(1,\frac{1}{2})$, *the distribution of* V *is* $\beta(\frac{1}{2},\frac{1}{2})$. *Let us rewrite* V *as*

$$V = \frac{W}{W+Z},$$

where W *and* Z *are independent, and* $W \sim \Gamma(\frac{1}{2},\frac{1}{2})$, *and* $Z \sim \Gamma(\frac{1}{2},\frac{r}{2})$. $\frac{Z}{r} \to 1$ *stochastically, thus* $rV \to W$ *in distribution, and thus* $\sqrt{r}\zeta_{1,r} \to \sqrt{UW}$. *The probability density functions for* $\sqrt{U}$ *and* $\sqrt{V}$ *are* $f_{\sqrt{U}}(x) = 2\varphi(x)$, *and* $f_{\sqrt{W}}(x) = xe^{-x^2/2}$. *Therefore,*

$$f_{\sqrt{UV}}(x) = \int_0^\infty f_{\sqrt{U}}(u) f_{\sqrt{W}}(\frac{x}{u})\frac{du}{u} = \int_0^\infty f_{\sqrt{W}}(u) f_{\sqrt{U}}(\frac{x}{u})\frac{du}{u},$$

so

$$f_{\sqrt{UV}}(x) = \frac{2}{\sqrt{2\pi}}\int_0^\infty \frac{x}{u^2} exp(-\frac{1}{2}(u^2+\frac{x^2}{u^2}))du =$$

$$\frac{2}{\sqrt{2\pi}}\int_0^\infty exp(-\frac{1}{2}(u^2+\frac{x^2}{u^2}))du =$$

$$\frac{1}{\sqrt{2\pi}}\int_0^\infty (\frac{x}{u^2}+1)exp(-\frac{1}{2}(u^2+\frac{x^2}{u^2}))du.$$

Substitute $t = u - \frac{x}{u}$. *Then* $dt = (\frac{x}{u^2}+1)du$ *and* $u^2 + \frac{x^2}{u^2} = t^2 + 2x$. *Therefore,*

$$f_{\sqrt{UW}}(x) = e^{-x}\int_{-\infty}^\infty \varphi(t)dt = e^{-x}.$$

□

CEU eTD Collection

4.3 Prüfer codes and uniform sampling

[E] The Prüfer code is a unique representation for trees, published by H. Prüfer in 1918 ([52]). A tree's Prüfer code can be easily obtained: consider a labeled tree T with n vertices, one of which nominated as the root. Remove $n-2$ of the vertices sequentially, such that at step i the leaf with the smallest label is removed, and, concurrently, the label of its parent is entered in $i-th$ position in the tree's Prüfer code. The resulting code will be unique, with length $n-2$.

Lemma 4. *Suppose $\{d_i\}(i=1,\ldots,n)$ is a sequence of positive integers, with $\sum_{i=1}^{n} d_i - 1 = n-2$, and let D be the distribution defined by $\{d_i\}$. Let $a_1,\ldots,a_n$ be independent, identically distributed variables, with $P(a_j=k)=\frac{d_k-1}{n-2}$. Then the tree that is Prüfer-encoded by $\{a_j\}$ is from the uniform distribution over the set of all labeled random trees whose Prüfer code is i.i.d. from D.*

Proof. *Observe that $\{a_j\}$ is a Prüfer code on a set of n vertices. Moreover, $\sum_{i=1}^{n} d_i - 1 = n-2$ implies $\sum_{i=1}^{n} d_i = 2(n-1)-2$, which means $\{d_i\}$ corresponds to the degree sequence of a tree on n vertices with one edge deleted.*

Suppose the tree corresponding to $\{a_j\}$ is T, with degree sequence $f_1,\ldots,f_n$. Then

$$L(T) = \prod_{i=1}^{n} \left(\frac{d_i-1}{n-2}\right)^{f_i-1}.$$

This formula only contains the degree sequence of the outcome, which implies the statement. □

Corollary. *If $\{a_j\}(j=1,\ldots,n-2)$ is a Prüfer code with degree sequence $\{f_i\}(i=1,\ldots,n)$, and π is a permutation on $n-2$ elements, then $\pi(a)$ corresponds to a tree sampled uniformly from the set of all labeled trees with degree sequence $\{f_i\}$.*

CEU eTD Collection

4.4 Measuring preferential attachment on trees

We have seen above that the set of permutations on the Prüfer code of a tree T correspond to a uniform mixing of all labeled trees on $\mu(T)$. In this section we will show how we can distinguish T from $\mu(T)$.

The statistic we will use is as follows. Suppose T is labeled with labeling l, and $root(T) = r$. For all of T's vertices i, let f_i be the family size of vertex i, as defined in section 4.1. Define statistic D as follows:

$$D = P(l(T) \text{ conservative }) =$$

$$\sum_{r \in V(T)} P(\text{labeling conservative} \mid root(T) = r) = \sum_{r \in V(T)} \prod_{i \in V(T)} \frac{1}{f_i}.$$

Since for any subtree T', the probability over all possible labelings that its root has the lowest label on the tree is exactly $\frac{1}{|V(T')|}$, the probability that the same holds for all of T's subtrees is $\prod_{i \in V(T)} \frac{1}{f_i}$, and the above formula follows.

Thus, D is a likelihood. Suppose now that T is a tree nascent from our model, and U is a Prüfer-mixed version of it, i.e. $U \in Unif(\mu(T))$. Then, by the Neyman-Pearson lemma, their likelihood ratio is a most powerful test. Thus, if D does not show a significant difference between T and U, we no longer need search in this direction.

[N] The table of results on the comparison between $D(T)$ and $D(U)$ is below. For notation, T will be from our model, U_1 and U_2 represent the minimal and maximal $D(U)$-values out of 250 Prüfer-mixes on T.

The test does show a significant difference already for relatively small trees. Therefore, a tree generated by our model does indeed stand out from its degree class.

This implies a tree nascent from this model is in fact distinguishable from its degree

CEU eTD Collection

Table 4.1: Prüfer mixing of a scale-free tree

$\lvert V(T)\rvert$	$\log(D(T))$	$\log(D(U_1))$	$\log(D(U_2))$
100	-40.53	-59.46	-39.83
500	-217.59	-315.69	-238.15
2000	-974.43	-1353.6	-1151.37
10000	-4825.33	-6463.91	-6068.73

sequence class, and therefore the degree sequence is insufficient in describing the model, and the reverse implication for 4.1 fails.

CEU eTD Collection

Chapter 5

General scale-free graphs

[**T**] The motivating problem for this chapter is the same as in the previous one: we wish to understand whether the reverse implication of 4.1 holds.

Albert and Barabási developed a model that managed to explain a power-law distribution outcome with exponent -3. Numerous models have since been introduced. In this chapter we will generalize the original model, in part to eliminate their weaknesses, and in part to demonstrate that it is not a difficult task.

5.1 Generalizations of the Albert-Barabási model

Since the Albert-Barabási model was introduced, numerous authors have worked on generalizing and redefining it. There seems to be general agreement that modeling it as a blown-up tree is very helpful in learning about some of its properties – see work by Bollobás, Durrett, Móri and others that we referred to earlier.

If a graph can be decomposed into a tree, we should get to a conclusion similar to that of the previous chapter. But will not employ tree decomposition, as Prüfer-mixing

CEU eTD Collection

those trees would break down the vertex order in the graph, and thus is inappropriate for our purposes. Instead, we will examine statistics that tell us information about a graph's preference structure, and show their relevance by comparing a graph against its degree sequence class. Analogous to the previous chapter, we will develop an approximate likelihood formula for these nascent graphs and show that they are indeed distinguishable.

5.1.1 Model I - The Albert-Barabási model

First, let us reformulate the model.

Let κ be any positive integer and G_0 an arbitrary labelled and directed graph, $n = |V(G)| \geq \kappa$. Let

$$W(G_0, \kappa) = \{U \subset G_0 : |U| = \kappa\}$$

For any $U \subset W$, let

$$\pi(U) = \prod_{u \in V(U)} d_u$$

Let $\Psi(G_0, \kappa) = \sum_{U \subset W} \pi(U)$.

On $W(G, \kappa)$, define the random variable X:

$$\forall U \subset W : P(X = U) = \frac{\pi(U)}{\Psi(G_0, \kappa)}$$

Now let $\tilde{G}$ be the graph that is created from G by adding a new vertex v to the vertex set of G and defining $\varepsilon_{v,u}$ as the indicator function for all $u \in V(G)$, where $\varepsilon_{v,u} = 1$ represents a directed edge from node v to u, and 0 represents no edge.

We will use the notation G_n for the graph we attain after the n-th step described as above.

This model is perhaps a good introduction, but has numerous weaknesses, notably:

CEU eTD Collection

- As the authors of the model point out, each new node connecting to the existing set through a constant set of edges is very restrictive.
- The scaling exponent is always 3.
- We hope to have a model with independent, rather than conditionally independent edges.

5.1.2 Model II - Independent dynamics

Let κ be any positive number and G an arbitrary labelled and directed graph, $|V(G)| \geq \kappa$ and $|E(G)| \geq \kappa$.

Now let $\tilde{G}$ be the graph that is created from G by adding a new vertex v to the vertex set of G and let $\varepsilon_{v,u}$, (for all $u \in V(G)$) independent conditioned on G with

$$P(\varepsilon_{v,u} = 1 \mid G) = \frac{\kappa d_u}{2|E(G)|}$$

This model no longer has the constraint that the outgoing degree of each node is constant. The edges, however, are still not independent, and the scaling exponent remains -3.

5.1.3 Model III - Independent dynamics with varied coefficients

Let $\{\beta_k\}$ be a sequence of positive numbers, $k = 1, \ldots, n$, and G_1 a graph with two vertices, 0 and 1, and one edge.

Now let G_{n+1} be the graph that is created from G_n by adding a new vertex v_{n+1} to the vertex set of G_n. and defining $\varepsilon_{n+1,u}$ as above. For $u = 0, \ldots, n$, $0 \leq p \leq 1$

$$P(\varepsilon_{n+1,u} = 1 \mid G_n) = min(p, \frac{\beta_{n+1} d_u}{2|E(G_n)|}).$$

CEU eTD Collection

The construction is very similar to Model II, but is far richer. For instance:

- $\beta_i = const$ yields a model II graph with power law distribution,
- $\beta_i = c(i-1)$, for $i = 2, \ldots,$ yields at step n a graph with $E(d_i) = const$ for all $i = 1, \ldots, n$.

This latter case resembles an Erdős-Rényi graph in the sense that it has identical expected degrees. But while an Erdős-Rényi graph may be sparse, this one will always be dense.

5.1.4 Model IV - Independent edges with exponent -3

Bollobás, Riordan, Spencer and Tusnády show in [12] that in the Albert-Barabási model, the expected degree of vertex i is

$$E(d_i) = \kappa\sqrt{\frac{t}{t_i}},$$

where t_i denotes the time point vertex i is connected to lower-index vertices.

If we measure time in identical, discrete intervals, we can paraphrase the above as

$$E(d_i) = \kappa\sqrt{\frac{n}{i}}.$$

This yields a scaling exponent of -3, as before.

Now, for the definition: Let κ be any positive number and G an arbitrary labelled and directed graph with n vertices. For $1 \leq i < j \leq n$ and $0 \leq p \leq 1$, let the random variables $\varepsilon_{j,i}$ be independent zero-ones representing a directed edge from vertex j to i, with

$$P(\varepsilon_{ji} = 1) = \min(p, \frac{\kappa}{2\sqrt{ij}}).$$

CEU eTD Collection

This model dealt with the weaknesses of the previous two, and allows for very easy computations. It has a weakness, however. To achieve κn as the expected total degree, we should have $P(e(j,i)=1)=\frac{\kappa}{2\sqrt{ij}}$. But that means for certain (κ,i,j) combinations we will get probabilities greater than 1, so p is used as a cutoff, which will interfere with the degrees of the leading vertices.

One very nice feature of this model is that it generates the graph as a diadic product, i.e. the probability of an edge between any pair of vertices depends exclusively on the degree sequence. In chapter 6 we will make use of this property.

5.1.5 Model IVa - Independent edges

Here we will introduce a model that will yield a scale-free distribution graph with an arbitrary exponent. This is just one possible model, aimed to demonstrate that the exponent -3 is by no means a strict limitation for an independent edge model. Cleary, the -3 exponent case must be a special case for the model.

Lemma 5. *Suppose G is a labeled graph with n vertices, with $d_1 \geq d_2 \geq \cdots \geq d_n$, and $d_i = \kappa(\frac{n}{i})^\alpha$, for some $\kappa > 0$ and $0 < \alpha < 1$. Then for an arbitrary $v \in V(G)$, $P(deg(v)=k) \sim k^{-\lambda}$, where $\lambda = \frac{\alpha+1}{\alpha}$.*

Proof.

$$P(deg(v)<k) = P(\kappa(\frac{n}{v})^\alpha < k) = P(v > \frac{n\kappa^{1/\alpha}}{k^{1/\alpha}}) = 1 - \frac{n\kappa^{1/\alpha}}{k^{1/\alpha}}.$$

We get the probability density through differentiation:

$$P(deg(v)=k) = \frac{n\kappa^{1/\alpha}}{\alpha k^{1+1/\alpha}} \sim k^{-\frac{\alpha+1}{\alpha}}.$$

□

Now, for the definition:

CEU eTD Collection

Let G be a labeled graph with n vertices, and $0 < \alpha < 1$. For $1 \leq i < j \leq n$ and $0 \leq p \leq 1$, let the $\varepsilon_{j,i}$ be independent zero-ones representing a directed edge from vertex j to i, with

$$P(\varepsilon_{ji} = 1) = \frac{\kappa\alpha}{i^{\alpha} j^{1-\alpha}}.$$

The expected degrees are:

$$E(d_i) = \sum_{j=1, j\neq i}^{n} P(\varepsilon_{j,i}) \approx \int_1^i \frac{\kappa\alpha}{j^{\alpha} i^{1-\alpha}} dj + \int_i^n \frac{\kappa\alpha}{i^{\alpha} j^{1-\alpha}} dj \approx$$

$$\frac{\kappa\alpha}{1-\alpha}(1 - \frac{1}{i^{1-\alpha}}) + \kappa(\frac{n}{i})^{\alpha} - \kappa.$$

For small values of i, the second term is dominant, for $i \sim n$, the $\frac{1}{i^{1-\alpha}}$ term can be neglected, so we can approximate

$$E(d_i) = \kappa\frac{2\alpha - 1}{1-\alpha} + \kappa(\frac{n}{i})^{\alpha}.$$

For $\alpha = \frac{1}{2}$, this is identical to Model IV, otherwise the degrees are going to be different from the expected $\kappa(\frac{n}{i})^{\alpha}$ for $i \approx n$.

5.2 Attaining uniform distribution

We have already considered the meaning of the term "uniform distribution" in chapter 3.3. In the present case, we will have to work on a labeling problem on an unlabeled graph. We will therefore define the uniform distribution as follows:

Let G be a labelled graph with n vertices, and $\mu(G)$ the class of graph such that

$$\forall H \in \mu(G) : \forall i = 1, \ldots, n, d_i(G) = d_i(H).$$

CEU eTD Collection

A uniform sample from the class of G will simply be a uniform sample from $\mu(G)$.

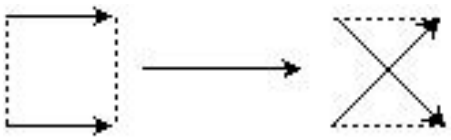
We saw in the previous chapter that a scale-free degree distribution does not imply preferential attachment on trees. To investigate this problem on richer graphs, we need to introduce an operation that does not modify a graph's degree sequence. This is the edge swap operation, and since it is widely known, we will only provide the definition for review.

[**E**]

Definition. *Let G be a graph with vertex set $V = \{v_1, \ldots, v_n\}$ and edge set $E = \{e_1, \ldots, e_m\}$. Consider two distinct edges $e_i = (v_1, v_2)$ and $e_j = (w_1, w_2)$. Then an edge swap on e_i and e_j is as follows:*

1. *If G is directed: we set $f_1 = (v_1, w_2)$ and $f_2 = (w_1, v_2)$. If $f_1, f_2 \notin E$, then e_i and e_j are deleted, and f_1, f_2 are added to the edge set.*

Figure 5.1: Directed swap



2. *If G is not directed: with probability $\frac{1}{2}$ each, we set ($f_1 = (v_1, w_2)$ and $f_2 = (w_1, v_2)$) or ($f_1 = (v_1, w_1)$ and $f_2 = (v_2, w_2)$). If $f_1, f_2 \notin E$, then e_i and e_j are deleted, and either f_1, f_2 are added to the edge set.*

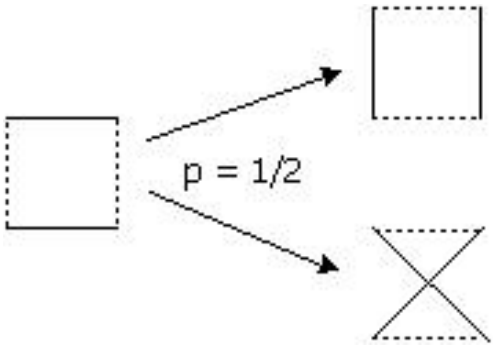
We defined swaps in a way that no double edges are created in the process.

Note: the term "edge swap" without specifically noting it is directed will mean nondirected for the rest of this paper.

Let us look at some swap properties. These are widely known, but we will provide our own proof for revision.

CEU eTD Collection

Figure 5.2: Non-directed swap



Lemma 6. *Consider a set V of n vertices, and an arbitrary graphic degree sequence $\{d_1, \ldots, d_n\}$. Let $G_1(V, E_1)$ and $G_2(V, E_2)$ be labeled graphs over V with corresponding edge sets E_1 and E_2, such that the degree sequence of both graphs is $\{d_i\}$. Then there exists a sequence of graphs $H_0, H_1, \ldots, H_{k+1}$, with $H_0 = G_1$ and $H_{k+1} = G_2$, with H_i exactly one swap away from its H_{i+1} for all $i = 0, 1, \ldots, k$, and the Hamming distance $d(H_i, G_2)$ monotone decreasing with i.*

Proof. *It might be helpful to think of a two-way process: we start off from G_1 and G_2, and in each step perform a swap on one of them, until we get two identical graphs. This is identical to the lemma's claim.*

Now, let $E = E_1 \bigcap E_2$. Color the edges in $E_1 \backslash E$ red, and those in $E_2 \backslash E$ blue.

We ought to make two observations.

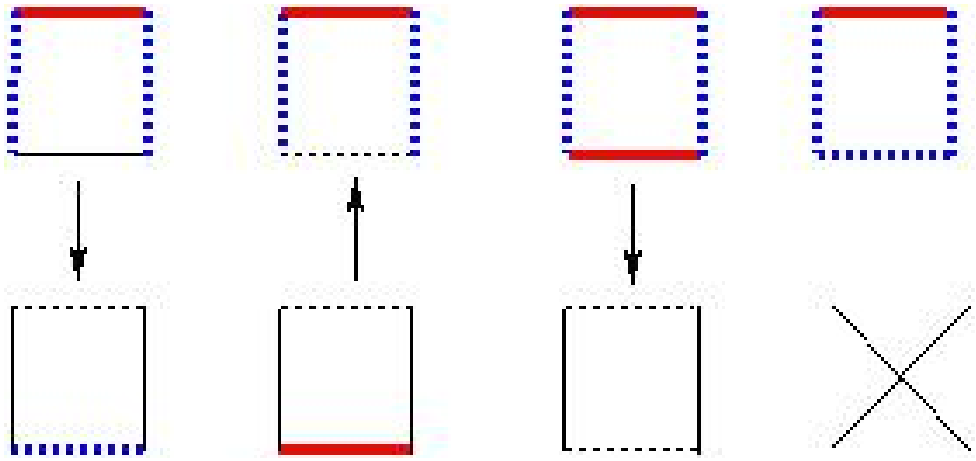
1. *For any vertex v_i, $deg_{red}(v_i) = deg_{blue}(v_i)$, since $deg(v_i) = d_i$ both in G_1 and G_2.*

2. *Given four vertices v_1, v_2, v_3, v_4 with (v_1, v_2) and (v_3, v_4) red, and (v_2, v_3) blue, a swap can be performed on these vertices unless (v_4, v_1) is red. If (v_4, v_1) is blue or not in $E_1 \cup E_2$, then the solution is automatic, since the red edges do exist in G_1, while the blue do not. If $(v_4, v_1) \in E$, a swap can now be performed backwards from G_2, since the red edges do not exist in G_2, but blue ones do. In all the above cases, a swap will*

CEU eTD Collection

decrease the Hamming distance. The only problematic case is then the ominous edge is red. See figure 5.3 for illustration.

Figure 5.3: Allowed swaps



Let H_i be the graph after $s_1, \ldots, s_i$ are performed on G_1, and J_i be the graph after $s_k, \ldots, s_{k-i+1}$ are performed on G_2. Clearly, $H_i = J_{k-i+1}$.

We will now show that we can always find an edge swap $H_i \to H_{i+1}$ or $H_j \to H_{j-1}$, such that the Hamming distance between the two working graphs decrease.

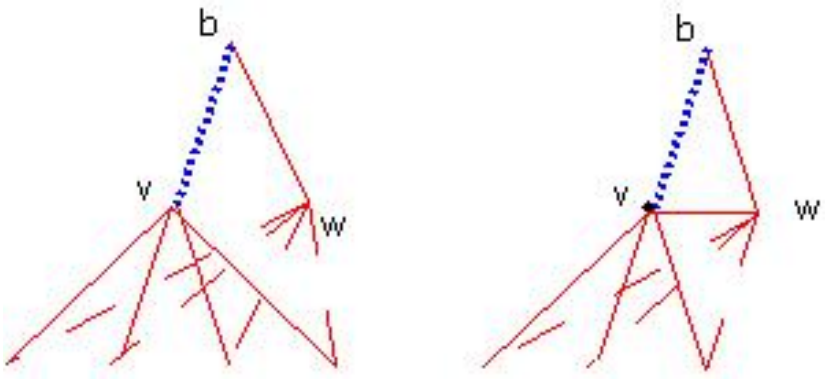
Suppose, by contradiction, that we cannot, and take a vertex v with a maximal red (and blue) degree d. Let b be a vertex with (b, v) blue, and let w be a red neighbor of b.

If w is not a red neighbor of v, then by the second observation, all of v's red neighbors must also be red neighbors of w, and thus $d_{red}(w) > d$, a contradiction. Or, if w is a red neighbor of v, then all of v's other red neighbors must also be red neighbors of w, plus v itself is a red neighbor, so $d_{red}(w) > d$, a contradiction. This completes the proof. □

The above lemma states that any two graphs with the same degree sequence can be transformed into each other via swaps. Wishful thinking whispers into our ear that if we could find a clever, randomized way of choosing swaps, we might be able to use this

CEU eTD Collection

Figure 5.4: Both cases lead to contradiction: $d_w > d_v$



operation to obtain uniform distribution over the class of graphs with an identical degree sequence.

Lemma 7. *Let $G(n, m)$ be a labeled directed graph, and $\mu(G)$ its degree sequence class. Call H_k the graph attained from G by k directed edge swaps. As $k \to \infty$, $H_k \subset Unif(\mu(G))$.*

Proof. *Let J be the metagraph of G with $V(J) = \mu(G)$, and $\forall u, v \in V(J)$, $(u, v) \in E(J)$ if and only if u and v are exactly one swap distant from each other. Our process is a random walk on the vertices of J along its edges.*

Clearly, the regularity of J is a sufficient condition for uniform mixing. J is not regular yet, but can be made so very easily. For all $u, v \in V(J)$, if $(u, v) \notin E(J)$, create a loop edge $u \to u$. In this construction, every vertex has degree $\binom{m}{2}$ in J.

For $u \in V(J)$, if the randomly chosen edge pair $e, f \in E(u)$ are swappable, we move from u to the neighbor marked by the swap, and stay otherwise. Every vertex in J has degree $\binom{m}{2}$, and every edge in it has probability $1/\binom{m}{2}$, and the statement follows. □

Lemma 8. *Let $G(n, m)$ be a labeled non-directed graph, and $\mu(G)$ its degree sequence class. Call H_k the graph attained from G by k non-directed edge swaps. As $k \to \infty$, $H_k \subset Unif(\mu(G))$.*

CEU eTD Collection

Proof. *The proof is analogous to the directed case, except here we will regularize the meta-graph to have degree* $2\binom{m}{2}$ *everywhere, with each edge having probability* $1/2\binom{m}{2}$*. This corresponds simply to the fact that for all* $u \in V(J)$*, there are two possible swaps that correspond to each edge pairing.* □

So we have shown that properly constructed edge swaps will indeed yield uniform mixing. Ideally, we would now need to show that the mixing time is fast enough. A new result from P.L. Erdős, I. Miklós and L. Soukup shows polynomial mixing time for a restricted class of graphs.

[O]

Theorem. *(Erdős, Miklós, Soukup) Let* G *be a semi-regular bipartite graph. Then the above-defined edge swaps will yield uniform distribution over* $\mu(G)$ *in polynomial time.*

The proof can be found in [26]. The general case is currently unsolved.

5.3 Not all degree distributions are equal

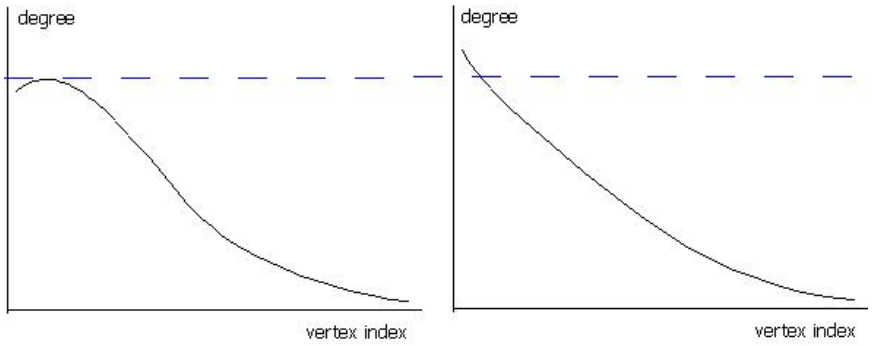
[T] We have defined several scale-free models, some sequential, some independent. There are several differences between them, of course. For example, in a plain AB-model graph, each vertex has a limited number of neighbors among the earlier vertices, but its last vertices have almost constant degree, unlike in the case of a, say, model III graph.

One more difference that is perhaps less obvious at first sight is in the degree distribution at the head. The expected degree for vertex i is a monotone decreasing function whose decline is steepest in the head. While independent edge models behave as expected in this regard, sequential models are different. The maximum degree node there is not the first node in the general case. See the figure.

Móri showed in a 2005 paper ([49]) that with probability 1, the identity of the highest degree node changes only a finite number of times, mainly while the graph is small. This is

CEU eTD Collection

Figure 5.5: The degree distribution of identically parametrized nascent model III vs. IV graphs with $\beta_k = const$ for model III



the reason behind the phenomenon. The same does not apply to independent edge models, and hence the difference in the head of the distributions.

5.4 Testing preferential attachment with the S metric

[**H**] We will now turn our attention to testing preferential attachment. The method will be to take a nascent scale-free graph G, and compare it with uniform samples G' from its degree sequence class. As we will see, the S metric can indeed distinguish G from G'. We will first provide the estimates, and then show some sample results.

5.4.1 S estimates: -3 exponent models

We have the task of calculating the expected value for the S metric on $G(n)$, as generated by our models. Observe that in terms of calculations, models I through IV are quite similar, therefore, we will only investigate model IV. Model III can be considered similar, as far as the β_i coefficients are constant, which we may assume for this exercise.

Observe that for $\kappa \geq \sqrt{8}$, it is formally possible that $\frac{\kappa}{2\sqrt{ij}} > 1$.

For $\kappa \leq \sqrt{8}$:

CEU eTD Collection

$$E(S(G)) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P((i,j) \in E(G)) d_i d_j = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\kappa^3 n}{2ij} \approx$$

$$\frac{1}{2} \int_1^n \int_1^n \frac{\kappa^3 n}{2xy} dxdy \approx \frac{\kappa^3 n \log^2 n}{4}$$

If $\kappa > \sqrt{8}$, then for indices $\leq \sqrt{\frac{\kappa}{2}}$, the probability of an edge can formally surpass 1. However, we should expect just an $O(\kappa)$ change in the total degree of the graph.

Indeed, simple calculations show that for a model IV graph G:

$$E(S(G)) = \frac{\kappa^3 n \log^2 n}{4} + O(n) \approx \frac{\kappa^3 n \log^2 n}{4}. \tag{5.1}$$

as expected.

5.4.2 Swap analysis on -3 exponent models

We saw the table of results: the value of S drops once a sufficient number of edges are swapped. Intuitively, this should mean that the probability of edges between degree-rich nodes are higher in our nascent graph G than from a graph from $\mu(G)$, and simulations confirm that. Let us now try to find an analytical explanation. Ideally, we should show the phenomenon through determining p_{uv} for any vertex pair $1 \leq u < v \leq n$, but our attempts were unsuccessful. Instead, we will provide an explanation based on heuristic approximations.

Suppose we have a graph $G(n, \kappa)$ nascent from a -3 exponent model. There are several questions we can ask:

- Which nodes' contribution to S will drop / grow?
- What is the expected effect of swaps on S?
- What is the expected value of S over $\mu(G)$?

CEU eTD Collection

Let us see these one by one. We will be looking at the non-directed case only. First, individual nodes' contribution.

Let us denote by $r_v(G)$ the neighborhood index or r in G, with

$$r_v(G) = \sum_{(v,j)\in E(G)} d_j.$$

Then a node v's contribution to the graph's S value is $d_v r_v(G)$.

To understand the contribution of individual vertices, we can simply investigate how the expected value for r_v is affected by a swap.

Proposition 3. *Suppose $G(n,m)$ is scale-free with $d_i \approx \kappa\sqrt{\frac{n}{i}}$ for all $i = 1,\dots,n$. Suppose the graph's edges are swapped as described before, forming G'. Then*

$$E(r_{u_1}(G') - r_{u_1}(G)) : \begin{cases} = 0 & \text{if } u_1 = \frac{4n}{(\log n)^2} \\ < 0 & \text{if } u_1 < \frac{4n}{(\log n)^2} \\ > 0 & \text{otherwise} \end{cases}$$

Proof. *Let us pick two distinct edges, with end vertices u_1, u_2 and u_3, u_4. A swap will remove these two edges and add either $\{(u_1,u_3),\ (u_2,u_4)\}$ or $\{(u_1,u_4),\ (u_2,u_3)\}$, if no double edges would be created by this step.*

To approximate the expected effect of swaps on S as a function of u_1, we need to compute

$$R(u_1) := E(r_{u_1}(G') - r_{u_1}(G)) = \frac{1}{3!}\frac{1}{2}\frac{1}{\kappa n}\int_1^n\int_1^n\int_1^n p_{u_1u_2}p_{u_3u_4}$$

$$[(1-p_{u_1u_3})(1-p_{u_2u_4})(d_{u_3}-d_{u_2}).$$

Now, $p_{vw} = \frac{d_v d_w}{2\kappa n}$. To avoid confusion with differentials, let us denote the degree of vertex v by t_v (i.e. $t_i = d_i$). We know that $\forall v = 1,\dots,n:\ t_v \approx \kappa\sqrt{\frac{n}{v}}$, which means

CEU eTD Collection

$v = \frac{\kappa^2 n}{{t_v}^2}$, *and thus*

$$dv = \frac{-2\kappa^2 n}{{t_v}^3} dt_v.$$

Furthermore, since both u_3 *and* u_4 *are integrated on the same interval, we do not have to process consider both possible swaps cases. Therefore, we can assume that the newly created edges should be* (u_1, u_3) *and* (u_2, u_4) *and leave the* $\frac{1}{2}$ *factor. Thus,*

$$R(u_1) = \frac{-8\kappa^5 n^2}{3!} \int_{\kappa\sqrt{n}}^{\kappa} \int_{\kappa\sqrt{n}}^{\kappa} \int_{\kappa\sqrt{n}}^{\kappa} \frac{t_1 t_2 t_3 t_4}{4\kappa^2 n^2} \left(1 - \frac{t_1 t_3}{2\kappa n}\right) \left(1 - \frac{t_2 t_4}{2\kappa n}\right) (t_3 - t_2) \frac{dt_2 dt_3 dt_4}{{t_2}^3 {t_3}^3 {t_4}^3}.$$

Due to the symmetry of t_3 *and* t_2*, this reduces to*

$$R(u_1) = \frac{-\kappa^2}{3!n} \int_{\kappa}^{\kappa\sqrt{n}} \int_{\kappa}^{\kappa\sqrt{n}} \int_{\kappa}^{\kappa\sqrt{n}} t_1 t_2 t_3 t_4 (t_1 t_3 + t_2 t_4)(t_3 - t_2) \frac{dt_2 dt_3 dt_4}{{t_2}^3 {t_3}^3 {t_4}^3} =$$

$$\frac{-\kappa^2}{3!n} \int_{\kappa}^{\kappa\sqrt{n}} \int_{\kappa}^{\kappa\sqrt{n}} \int_{\kappa}^{\kappa\sqrt{n}} \left(\frac{{t_1}^2}{{t_2}^2 {t_4}^2} - \frac{{t_1}^2}{t_2 t_3 {t_4}^2} + \frac{t_1}{t_2 t_3 t_4} - \frac{t_1}{{t_3}^2 t_4}\right) dt_2 dt_3 dt_4.$$

Substituting $a = \kappa\left(\sqrt{n} - 1\right)$, $b = \frac{1}{\kappa}\left(1 - \sqrt{\frac{1}{n}}\right)$, *and* $c = \log\sqrt{n}$, *this computes to*

$$R(u_1) = -\frac{\kappa^2 t_1}{3!n}(bt_1 - c)(ab - c^2).$$

The only factor that can be negative in the above expression is $bt_1 - c^2$*.* R *will change its sign where this monotone expression equals* 0*. Solving it yields*

$$\frac{\sqrt{n} - 1}{\sqrt{u_1}} - \log\sqrt{n} = 0 \Rightarrow$$

$$u_1 \approx \frac{4n}{(\log n)^2},$$

with $u < u_1 \Rightarrow R < 0$*, and* $u > u_1 \Rightarrow R > 0$*, as stated.*

□

CEU eTD Collection

.

Corollary. *Suppose $G(n,m)$ is scale-free with $d_i \approx \kappa\sqrt{\frac{n}{i}}$ for all $i = 1,\ldots,n$. Suppose the graph's edges are swapped as described before, forming G'. Then*

$$E(S(G') - S(G)) < 0.$$

Proof. *We will use the notation from the above proposition. Observe that in general for graph $H(n)$ we have*

$$S(H) = \sum_{i=1}^{n} \frac{1}{2} d_i r_i(H),$$

thus for graphs G and G':

$$E(S(G') - S(G)) = \frac{1}{2} \sum_{u_1=1}^{n} d_{u_1} R(u_1) \approx$$

$$-\kappa^2 n \int_{\kappa}^{\kappa\sqrt{n}} t_1 \frac{\kappa^2 t_1}{3!n}(bt_1 - c)(ab - c^2)\frac{dt_1}{{t_1}^3} = -\frac{\kappa^4}{3!}(ab - c^2)^2 < 0.$$

□

.

Corollary. *For nascent scale-free graph $G(n,\kappa)$, let $G' \in \mu(G)$. Then*

$$E(S(G')) \approx \frac{\kappa^3 n}{4}\left((\log n)^2 - \frac{4\kappa}{3}\right) \approx \frac{\kappa^3 n(\log n)^2}{4}, \tag{5.2}$$

asymptotically identical to the nascent value.

Let us now compute S for model IVa graphs.

CEU eTD Collection

5.4.3 S estimates: Model IVa

We defined in model IVa p_{ij} $(i < j)$ as:

$$p_{ij} = \frac{\kappa\alpha}{i^{\alpha} j^{1-\alpha}}.$$

Thus, for $G \subset M_{IVa}(n)$:

$$E(S(G)) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{ij} d_i d_j \approx \int_1^n \int_i^n \frac{\kappa^3 \alpha n^{2\alpha}}{i^{2\alpha} j} dj di =$$

$$\kappa^3 \alpha n^{2\alpha} \int_1^n \frac{\log n - \log i}{i^{2\alpha}}$$

We know the result for $\alpha = \frac{1}{2}$. The two other cases:

Case 1: $0 < \alpha < \frac{1}{2}$.

$$E(S(G)) \approx \frac{\kappa^3 \alpha}{(1-2\alpha)^2} n,$$

Case 2: $\alpha = \frac{1}{2}$ (also seen in previous section).

$$E(S(G)) \approx \frac{\kappa^3}{4} n \log^2 n,$$

Case 3: $\frac{1}{2} < \alpha < 1$.

$$E(S(G)) \approx \frac{\kappa^3 \alpha}{2\alpha - 1} n^{2\alpha} \log n.$$

5.4.4 The value of S_{max}

[E] We have just seen what S values our models are expected to generate. How about S_{max}?

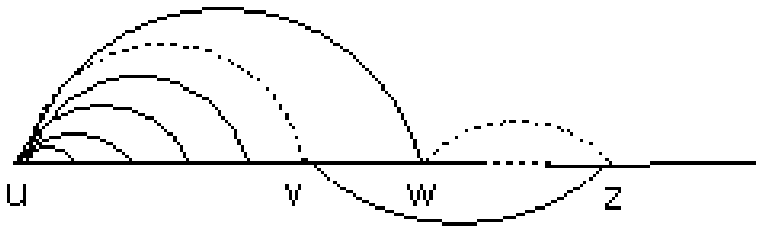
Lemma 9. *Let G be a graph with n vertices, labeled such that $d_1 \geq d_2 \geq \cdots \geq d_n$. If $\exists$ $v, w \in V(G)$ with $1 = u < v < w$ such that $(u, v) \notin E(G)$, but $(u, w) \in E(G)$, then there*

CEU eTD Collection

exists edge swap $G \to G'$, such that $S(G') \geq S(G)$, with equality holding only if $d_v = d_w$.

Proof. *Since $d_v \geq d_w$ and u is not a neighbor of v, there must exists some node z such that $(v, z) \in E(G)$, but $(w, z) \notin E(G)$.*

Figure 5.6: Proof illustration



Let G' be the graph attained from G by swapping $\{(u, w), (v, z)\}$ with $\{(u, v), (w, z)\}$. Since $u = 1$, it is true regardless of z that

$$S(G') - S(G) = d_u d_v + d_w d_z - (d_u d_w + d_v d_z) = (d_u - d_z)(d_v - d_w) \geq 0,$$

with equality only in case $d_v = d_w$. □

Corollary. *Let G be a graph with n vertices. There is a greedy algorithm that will locally maximize S through edge swaps.*

Proof. *We will prove the existence of the algorithm by construction. Let G be a graph as above, with $d_1 \geq \cdots \geq d_n$. Let G_i be a subgraph of G generated by $v_i, \ldots, v_n$, and let the expression $d_k(G_j)$ refer to the degree of v_k in G_j. Using the above proof's notation: for step i of the algorithm, set $u = i$, and then by the above lemma we can make sure u is connected to vertices $i + 1, \ldots, i + d_i(G_i)$.*

Observe that since for all vertices $1, \ldots, i - 1$ their neighbors are also chosen greedily, vertex z will definitely have index $> u$, thus step i does not interfere with any of the earlier steps. □

CEU eTD Collection

Theorem 2. *Let $\{d_i\}$ be a graphic degree sequence ($i \in \{1, \ldots, n\}$), with $d_i = \lfloor \kappa(\frac{n}{i})^\alpha \rfloor$. Let G be a graph with this degree sequence that maximizes S. Then $S(G) = \Omega(n^{4\alpha/\alpha+1})$.*

Proof. *We want to provide a lower estimate for $S(G)$ by construction. Let H be an empty graph with n points. Connect v_1 to $v_2, \ldots, v_{d_1+1}$. It can be done, as long as $d_{d_1+1} \geq 1$. Then connect v_2 to $v_3, \ldots, v_{d_2+1}$, which can be done as long as $d_{d_2+1} \geq 2$, and so on. For small values of k, the k-th vertex will have $k-1$ neighbors with index $< k$, and $d_k - k$ with index $> k$, so its highest index neighbor will be vertex $d_k + 1$. The process will terminate when $d_k \leq k$. The rest of the edges can be filled in freely: since the degree-richest nodes are heavily interconnected, the rest of the graph will be graphic (see considerations in proof of Erdős-Gallai theorem).*

Solving

$$d_k = k,$$

we get

$$k = \kappa(\frac{n}{l})^\alpha,$$

consequently

$$l = \sqrt[\alpha+1]{\kappa n^\alpha}.$$

That is, a complete subgraph of size $O(n^{\alpha/\alpha+1})$. Simple calculations yield that the value of S on this complete subgraph is $\Theta(n^{4\alpha/\alpha+1})$, which proves the statement. □

This, together with the S metric calculations for model IVa allow us to conclude that

Corollary. *For a nascent scale-free graph G with any $0 < \alpha < 1$, $\lim_{n\to\infty} \frac{S(G)}{S_{max}(\mu(G))} = 0$, and thus Li et al's claim in [42] that this ratio is close to 1 is incorrect.*

CEU eTD Collection

5.4.5 Results - non-directed case

[N] Our estimates for nascent the S value in (5.1) and its degree sequence class in (5.2) are asymptotically identical, but S is lower by some smaller term. Thus, simple linear histograms are very convenient for showing our results.

Table 5.1: S metric values in histogram, nondirected graph, $5|E(G)|$ edge swaps. For model IVa, $\alpha = 0.67$. The location of the nascent graph is signified by parentheses

Model	n	κ	histogram
I	500	5	1 0 0 (1) 0 0 1 1 4 4 1 3 4 4 2 5 4 3 2 2 1 1 1 2 3
I	500	20	1 4 4 4 9 11 9 4 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 (1)
I	1500	20	3 22 17 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 (1)
I	2000	30	13 31 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 (1)
IVa	500	5	1 1 0 0 4 3 1 2 4 3 5 4 2 0 4 (4) 3 2 1 0 2 2 2 0 1
IVa	500	20	1 0 1 0 0 1 4 3 3 1 6 7 5 5 4 3 2 2 1 1 0 0 0 0 (1)
IVa	1500	20	1 0 0 2 1 4 4 3 5 5 6 6 4 5 1 0 1 0 1 1 0 0 0 0 (1)
IVa	2000	30	1 0 0 5 3 4 5 5 4 9 5 4 2 2 1 0 0 0 0 0 0 0 0 0 (1)
III	500	5	1 0 2 0 5 2 3 1 3 4 1 4 1 1 4 2 (2) 2 3 3 4 1 1 0 1
III	500	20	4 1 12 10 7 4 8 3 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 (1)
III	1500	20	2 5 10 9 13 9 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 (1)
III	2000	30	8 7 20 12 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 (1)

This is just a short illustration of the phenomenon. We found in general that for smaller graphs and low edge densities, S did not necessarily decrease after a swapping the edges well, but the phenomenon became clear and reproducable as the graph's size and the edge density grew for all models.

5.4.6 Results - directed case

Let us see the same directed scale-free graphs. If the graph's edges are directed according to the logic our models would suggest – say, from later to earlier vertices –, then S drops more significantly and already at lower sizes and densities than in the undirected case. Remember that the distribution of the incoming and outgoing degrees are significantly

CEU eTD Collection

different, the outgoing degrees' distribution is roughly constant. So edges between high-degree vertices are far more likely to be broken up than reinstated.

The table below illustrates that S does drop more sharply in the directed case.

Table 5.2: S metric values in histogram, directed graph, $5|E(G)|$ edge swaps. The location of the nascent graph is signified by brackets

Model	n	κ	histogram
III	500	5	1 0 4 4 3 12 5 9 3 4 3 1 1 0 0 0 0 0 0 0 0 0 0 (1)
III	500	20	10 26 9 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 (1)
III	1500	20	14 32 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 (1)
III	2000	30	42 8 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 (1)

In conclusion to this section, we have seen both analytically and experimentally that the S metric distinguishes nascent scale-free graphs from their degree sequence class.

5.5 An urn-based likelihood model

[T] We have seen that the S metric does a relatively good job at identifying preferential attachment, but we know nothing about its power. One would hope to find a test that is easier to grasp, even if it not of maximum power.

Consider the following urn representation for a sequentially built preferential attachment graph.

Suppose the starting graph will have just two vertices, 0 and 1, and one edge. Starting with $k = 2$, vertex k is connected to the graph at time k, with $g_k(k)$ outgoing edges and $h_k(k) = 0$ incoming edges. As the graph grows, h_k might increase. The graph is not meant to be directed, we only keep a sense of direction for our calculations - the degree of vertex k at any time τ is $d_k(\tau) = g_k(\tau) + h_k(\tau)$.

Denote by f the function for the preferential attachment, and consider what happens when we add vertex t (at time t). The probability of choosing k as a neighbor is exactly

CEU eTD Collection

$\frac{f(d_k(t))}{\sum_{i=0}^{t} f(d_k(i))} =: q(t,k)$.

Vertex t is meant to have $g_t(t)$ neighbors from the set of vertices with $1, \ldots, t-1$. We will choose them independently. If k is chosen more than once, we discard our pickings, and select another group of $g_t(t)$ vertices.

Since vertex t is to have $g_t(t)$ neighbors at entry, the probability that none will connect to k is $M = q(t,k)^{g_t(t)}$, and

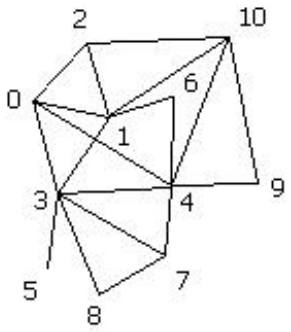
$$P((t,c) \in E(G(t))) = 1 - q(t,k)^{g_t(t)}.$$

For an explicit urn representation:

For vertex k, consider urns $A(k)$ and $B(k)$. $A(k,t)$ contains the degree number of vertex k, while urn $B(k,t)$ contains all other degrees of the graph (i.e. $= 2|E(G(t))| - A(k,t)$). To add vertex $t+1$, we draw $g_{t+1}(t+1)$ times from the urns according to f. If we picked from A more than once, we discard our choices and try again. Finally, we increase A by i and B by $2g_{t+1}(t+1) - i$, where i is 1 if we have picked a ball from A, 0 otherwise. This translates to a binary model with balls from A represented by 1 bits, and balls from B by 0-s.

The figure below might help our understanding.

Figure 5.7: An example graph



CEU eTD Collection

The graph in the above figure has $|V(G)| = 11$, $|E(G)| = 19$. Let us first see the urn representation of, say, vertex 4 for an illustration. All other vertices except 0 have similar representations of length $2|E(G)|$.

Table 5.3: Urn representation for vertex 4 in the above graph

code	00	0000	0000	0011	00	0100	0100	0000	10	00100000
neighbor	1	2	3	self	5	6	7	8	9	10

For graph $G(n)$ ($V(G) = \{v_0, \ldots, v_n\}$), we will call the n x $|2E(G)|$ matrix thus composed "the urn matrix for G".

Let us formalize the above, with a generalization. Suppose we hypothesize a preferential attachment model for graph $G(n, m)$, such that the model's preference relation for the neighbors of vertex τ from the vertex set $i = 1, \ldots, \tau - 1$ is given by

$$P(\varepsilon_{i\tau} = 1) = 1 - (\frac{f(h_i(\tau))}{\sum_{j=1}^{\tau-1} f(h_j(\tau))})^{g_t(\tau)},$$

for some f preference function.

Then the likelihood for the graph is

$$L(G) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} P(\varepsilon_{ij} = 1)^{\varepsilon_{ij}} (1 - P(\varepsilon_{ij} = 1))^{1-\varepsilon_{ij}}.$$

This is a quasi-likelihood value for G. The reason it cannot be called a proper likelihood function is that there was a little bit of cheating in deriving the formula. We did not allow multiple pickings from urn A in one round, but we did for urn B. This does not change the result very much in practice, but we ought to be aware of it.

Our experience with the S metric shows consistenly that these nascent graphs stand out sharply from their degree sequence class. This implies that the $0-1$ urn sequence assigned

CEU eTD Collection

to each vertex is close to the maximum likelihood sequence, for a well-chosen f. We saw in section 3.5 how to make some educated guesses about f. For a good estimate for f and a relatively good guess for the original vertex order, each line of the urn matrix should be close to its maximum likelihood. Thus, can search for f by making perturbations on the lines of the urn matrix.

5.5.1 Results - sequential case

[N] We can test the quasi-likelihood formula we derived above on a Model III graph. First, by making perturbations on the urn matrix as described above, to see how the overall likelihood changes. And second, by examining the same under edge swaps.

For notation, say $M(G)$ is the urn matrix as derived above, $M'(G)$ is a perturbation of the matrix, while G' is a perturbation on G itself, namely edge swaps. $L(G)$ and $L(M(G))$ are two ways to notate the same thing, but since we are comparing two different types of perturbations, we will use them both.

Table 5.4: Urn quasi-likelihood, model III with $\beta = const$, 25 times $5|E(G)|$ matrix perturbations

$\|V(G)\|$	$2\frac{\|E(G)\|}{\|V(G)\|}$	$\log L(M_G)$	$min\{\log L(M'_G\}$	$max\{\log L(M'_G)\}$
500	5	-15529.693	-9265.851	-8329.518
500	20	-62344.23	-49018.12	-47856.084
1500	20	-222063.139	-126231.107	-114772.315
2000	30	-471715.817	-266005.082	-244089.355

The same under edge swaps:

These results confirm that the preference structure is different in the nascent graph than over its degree class. The data obviously shows that the urn matrix perturbations affect L much more spectacularly than edge swaps. This seems to suggest that edge swaps have limited implications on the preference structure. That is reasonable, since the degree

CEU eTD Collection

Table 5.5: Urn quasi-likelihood, model III with $\beta = const$, 50 times $5|E(G)|$ edge swaps

$\|V(G)\|$	$2\frac{\|E(G)\|}{\|V(G)\|}$	$\log L(M_G)$	$min\{\log L(M_{G'}\}$	$max\{\log L(M_{G'}\}$
500	5	-15866.891	-13122.426	-12198.573
500	20	-59664.732	-58153.909	-46070.313
1500	20	-224914.946	-177270.854	-173037.467
2000	30	-463474.606	-368972.010	-362138.322

sequence is a strong constraint on the graph. However, this matter needs to be investigated further, we will consider it an open problem for this publication.

5.5.2 Results - nonsequential case

How about non-sequentially built graphs? Let us see the same for a model IVa graph with $\alpha = 0.67$. Note that model IVa has the same simple preferential attachment function $f(d) = cd$.

Table 5.6: Urn quasi-likelihood, model IVa, 25 times $5|E(G)|$ matrix perturbations

$\|V(G)\|$	$2\frac{\|E(G)\|}{\|V(G)\|}$	$\log L(M_G)$	$min\{\log L(M'_G)\}$	$max\{\log L(M'_G)\}$
500	5	-8432.755	-4897.828	-3986.803
500	20	-30741.212	-17647.714	-16424.394
1500	20	-108165.475	-61532.483	-56539.051
2000	30	-229222.397	-129596.355	-117620.626

CEU eTD Collection

Edge swaps:

Table 5.7: Urn quasi-likelihood, model IVa, 50 times $5|E(G)|$ edge swaps

$\|V(G)\|$	$2\frac{\|E(G)\|}{\|V(G)\|}$	$\log L(M_G)$	$min\{\log L(M_{G'})\}$	$max\{\log L(M_{G'})\}$
500	5	-7873.848	-6399.312	-5796.338
500	20	-30794.981	-25163.67	-23745.459
1500	20	-108844.844	-86663.854	-84546.459
2000	30	-227471.08	-179653.628	-175017.48

The results are very similar to the ones we have just seen. This is confirmation that

the model is a good approximation for a sequential model.

5.5.3 Results - directed case

What happens if the graph is directed? We have seen with the S metric that the swap changes the preference structure more than in the non-directed case. We expect to see the same phenomenon here. Note that we do not have to examine the urn matrix perturbations, since that will be the same regardless of direction.

Table 5.8: Urn quasi-likelihood, model III with $\beta = const$, 50 times $5|E(G)|$ edge swaps

$\|V(G)\|$	$2\frac{\|E(G)\|}{\|V(G)\|}$	$\log L(M_G)$	$min\{\log L(M_{G'})\}$	$max\{\log L(M_{G'})\}$
500	5	-14936.768	-13096.573	-10554.247
500	20	-63585.505	-55455.498	-51094.704
1500	20	-224021.59	-192885.545	-180562.461
2000	30	-465479.328	-401647.686	-382838.769

The results show a marked difference, although it is somewhat surprising that this difference is in general less than in the non-directed case, as it is contrary to our experience with the S metric. This will need further investigation.

5.5.4 Results - β_i and the preference function varied

A natural next step for the experiments was to vary β_i and the preference function, and look for any changes in the results. Surprisingly, nothing seemed to change – the above quasi-likelihood model with a linear preference function was very resilient even when the actual nascent preference functions was far from linear. Not to mention β_i. This will need to be revisited, but without proper analysis, the phenomenon does not merit more than a paragraph, which ends here.

CEU eTD Collection

5.6 Summary

[**T**] We have seen in this chapter that the reverse implication for 4.1 fail in the general scale-free case as well. One noteworthy phenomenon we found was that swapping does not have as dramatic an effect on a graph's preference structure as perturbing the urn model bit sequences. This is not very surprising by itself, but has far-reaching implications. Namely, it suggests that while the degree sequence is inadequate for fully understanding a graph, there is still a lower bound for the preference structure of the graph.

CEU eTD Collection

Chapter 6

The uniform model

[E] Consider a random graph G with n vertices, with independent edge probablities $p_{ij} = P((i,j) \in E(G))$. The likelihood of a particular graph in this model is

$$L(G) = \prod_{(i,j)\in E(G)} p_{ij} \prod_{(i,j)\notin E(G)} (1 - p_{ij}).$$

Let ε_{ij} be the indicator of an edge between i and j, i.e. $\varepsilon_{ij} = 1$ if (i,j) is an edge, $\varepsilon_{ij} = 0$ otherwise.

The above expression can be rewritten as

$$L(G) = \prod_{i,j\in V(G), i<j} (1 - p_{ij}) \prod_{\varepsilon_{ij}=1} \frac{p_{ij}}{1 - p_{ij}}.$$

Let $q_{ij} = \frac{p_{ij}}{1-p_{ij}}$. Model IV expressed p_{ij} as a diadic product, now, in similar vein, say that $q_{ij} = \alpha_i \alpha_j$ for some $\{\alpha_k\}, k = 1, \ldots, n$.

Then

$$L(G) = C \prod_{\varepsilon_{ij}=1} q_{ij} = C \prod_{\varepsilon_{ij}=1} \alpha_i \alpha_j = C \prod_{i=1}^{n} {\alpha_i}^{d_i}.$$

CEU eTD Collection

So the likelihood of the graph depends solely on the degree sequence. Or, identically put, the degree sequence is a sufficient statistic. This chapter will investigate a number of properties for this model. Many of these results are published in parallel in [21].

Now, for the formal definition for this model V:

Suppose n is a positive integer, and $\alpha_1, \ldots, \alpha_n$ are arbitrary positive numbers. The probability of an edge between vertices $1 \leq i < j \leq n$ is

$$p_{ij} = \frac{\alpha_i \alpha_j}{1 + \alpha_i \alpha_j},$$

all independent.

By merit of the likelihood depending exclusively on the degree sequence, this model generates graphs with uniform probability over the class of graphs with expected degree sequence $E(d_i) = \sum_{i \neq j} \frac{\alpha_i \alpha_j}{1+\alpha_i \alpha_j}$.

Furthermore, the model eliminates the rather inconvenient weakness of models IV and IVa that for certain combinations (κ, i, j) those could formally yield probabilities greater than 1.

The price is that there is no explicit formula for α_i in general. We can, however give an iterative approximation. First, consider the likelihood equation:

$$L(G) = \prod_{i<j} (1 - p_{ij}) \prod_{\varepsilon_{ij}=1} q_{ij} = \frac{\prod_{i=1}^{n} {\alpha_i}^{d_i}}{\prod_{i<j} (1 + \alpha_i \alpha_j)}.$$

Then $\log L = \sum_{i=1}^{n} d_i \log \alpha_i - \sum_{i<j} \log(1 + \alpha_i \alpha_j)$, and the maximum likelihood solution over $\{\alpha_i\}$ is at $\frac{\partial \log L}{\partial \alpha_i} = 0$ for $i = 1, \ldots, n$. Namely,

$$\frac{d_i}{\alpha_i} - \sum_{i \neq j} \frac{\alpha_j}{1 + \alpha_i \alpha_j} = 0, \text{ for } 1 \leq i \leq n.$$

CEU eTD Collection

This could be solved by the iteration

$$\tilde{\alpha_i} = \frac{d_i}{\sum_{i \neq j} \frac{\alpha_j}{1+\alpha_i \alpha_j}}. \tag{6.1}$$

Let us examine which $\{d_i\}$ sequences allow for a solution for the above iteration is solvable.

6.1 The graphs whose maximum likelihood equation can be solved

We have just introduced a very general model. We need to do some work in order to understand the set of graphs that allow us to solve the maximum likelihood equation, and have a meaning for the above iteration.

First of all, the Erdős-Gallai theorem ([28]) states that the necessary and sufficient condition that there exists a graph for degree sequence $\{d_i\}$, $i = 1, \ldots, n$:

$$\sum_{i=1}^{k} d_i \leq k(k-1) + \sum_{i=k+1}^{n} min(k, d_i)$$

for all $1 \leq k \leq n-1$.

Lemma 10. *Let G be a graph with n vertices and degree sequence $d_1, \ldots, d_n$. For all $i = 1 \ldots, n$, let $f_i = E(d_i) = \sum_{j \neq i} p_{ij}$. Then the set of n-dimensional vectors $\{f_1, \ldots, f_n\}$ is convex.*

Proof. *Observe that the set of p_{ij}-s ($i, j \in \{1, \ldots, n\}$) are contained in the $\binom{n}{2}$-dimensional*

CEU eTD Collection

unit cube. For a fixed $k \in \{1, \ldots, n\}$ and $u > k$, define

$$r_u = \begin{cases} 1 & \text{if } k < f_u, \\ 0 & \text{otherwise} \end{cases}$$

Then we can rewrite the Erdős-Gallai inequality for vertex k as

$$\sum_{i=1}^{k} f_i \leq k(k-1) + \sum_{i=k+1}^{n} (r_u k + (1 - r_u) f_i),$$

which is linear regardless of r_u, just as the expected value operation. The same holds for all $n - k$ inequalities, regardless of k, so overall we are performing a linear transformation on a convex set, which will result in a convex set, as stated. □

Let C denote the set of expected degree sequences for all graph realizations over n vertices. By the above, C is convex. We will now extend the Erdős-Gallai theorem over C.

Lemma 11. *Let P be an $n \times n$ symmetric matrix with $0 \leq p_{ij} \leq 1$ for all $i \neq j$, and 0 entries on the diagonal. Let $f_i = \sum_j p_{ij}$. Then $\forall k = 1, \ldots, n$*

$$\sum_{i=1}^{k} f_i \leq k(k-1) + \sum_{j=k+1}^{n} \min(k, f_j).$$

Proof. *Fix k and denote $\sigma_j = \sum_{i=1}^{k} p_{ij}$. Then for $j \leq k$, we have $\sigma_j \leq k - 1$, and for $j > k$, we have on the one hand $\sigma_j \leq k$, and on the other $\sigma_j \leq f_j$.* □

We know that the boundary of C is marked by those matrices that have some 0 or 1 entries outside the diagonal. A solution may not exist on the boundary, but it does if $\underline{f}$ is from the interior of C, as we will now show.

Theorem 3. *Let $\underline{f}$ be an interior point of C. Let $\nu(\underline{f})$ be the set of $n \times n$ symmetric probability matrices with $f_i = \sum_{j=1}^{n} p_{ij}$. Then the random model that maximizes the graph's*

CEU eTD Collection

binary entropy over the edge probabilities p_{ij}, given expected degrees $f_1, \ldots, f_n$ is model V, as defined above.

Proof. *For any matrix $P \in \nu(\underline{f})$, the graph generated from P will have an edge (i, j) with probability p_{ij}, or no edge with probability $1 - p_{ij}$.*

Consider the binary entropy function:

$$S = -\sum_{1 \le i < j \le n} p_{ij} \log p_{ij} + (1 - p_{ij}) \log(1 - p_{ij}).$$

Clearly $S > 0$, so it is bounded below. It will be maximized by a model that contains the structure minimally induced by the conditions on $\underline{f}$.

To maximize it, we will use the Lagrange multiplier method, which states that $\exists\{K_i\}$ such that

$$\tilde{S} = S + \sum_{i=1}^{n} K_i (f_i - \sum_{j \ne i} p_{ij}) \to \max.$$

To achieve that, we will have to solve for all $i \ne j$

$$\frac{\partial \tilde{S}}{\partial p_{ij}} = -1 - \log p_{ij} + 1 + \log(1 - p_{ij}) - K_i - K_j = 0.$$

This leads to

$$\log \frac{p_{ij}}{1 - p_{ij}} = -(K_i + K_j).$$

Substituting for all i: $K_i = -\log \alpha_i$, as in model V, we get

$$\frac{p_{ij}}{1 - p_{ij}} = \alpha_i \alpha_j,$$

exactly as defined by the model. □

We do not know what happens on the boundary of C, it remains an open question.

CEU eTD Collection

6.2 When will the iteration converge?

[O] We have seen some sufficient conditions for the existence of the maximum likelihood equation. We should now study if this is sufficient for solving the iteration on the coefficient sequence.

Thanks for the solution to T. F. Móri.

Lemma 12. *Suppose $G(n,m)$ has degree sequence $d_1,\dots,d_n$. If the model V maximum likelihood equation has a solution for $\{d_i\}$, then the iteration*

$$\tilde{\alpha_i} = \frac{d_i}{\sum_{i\neq j}\frac{\alpha_j}{1+\alpha_i\alpha_j}}.$$

will converge.

Proof. *First, we may assume that for all $1 < i \leq n$, $0 < d_i < n-1$. If $d_i = 0$, then*

$$\frac{\alpha_i^{d_i}}{\prod_{i\neq j}(1+\alpha_i\alpha_j)}$$

is maximized by $\alpha_i = 0$. We may ignore this case. If, on the other hand,$d_i = m-1$, then the above can be rewritten as

$$\prod_{i\neq j}\frac{\alpha_i}{1+\alpha_i\alpha_j},$$

which is monotone in α_i, and thus is maximized by $a_i = \infty$ (i.e. $p_{ij} = 1$ for all $i \neq j$). We may ignore this case as well.

Let us introduce a metric on $\mathbb{R}_+^n$:

$$\rho(\underline{\alpha},\underline{\beta}) = \max_{1\geq i\geq n} |\log\alpha_i - \log\beta_i|.$$

We will show that $\rho(\underline{\tilde{\alpha}},\underline{\tilde{\beta}}) \leq \rho(\underline{\alpha},\underline{\beta})$. In fact, for $n \geq 3$, the distance will certainly decrease after two iterative steps.

CEU eTD Collection

Let $c_i = \frac{\alpha_i}{\beta_i}$, *and* $\tilde{c_i} = \frac{\tilde{\alpha_i}}{\tilde{\beta_i}}$. *Suppose that* $\forall 1 \leq i \leq n$, $\frac{1}{K} \leq c_i \leq K$. *Then*

$$\tilde{c_i} = \frac{\sum_{j \neq i} \frac{\beta_j}{1+\beta_i\beta_j}}{\sum_{j \neq i} \frac{\beta_j c_j}{1+\beta_i\beta_j c_i c_j}} \leq^{(1)} \max_{j \neq i} \frac{1+\beta_i\beta_j c_i c_j}{c_j(1+\beta_i\beta_j)} = c_i \max_{j \neq i} \frac{\frac{1}{c_i c_j} + \beta_i\beta_j}{1+\beta_i\beta_j} \leq^{(2)}$$

$$c_i \max_{j \neq i} \max(\frac{1}{c_i c_j}, 1) = \max(\frac{1}{\min_{j \neq i} c_j}, c_i) \leq^{(3)} K$$

Conditions for equality:

- *in (2), for the* j *value that corresponds to the maximum,* $c_i c_j = 1$,
- *in (3),* $c_i = K$, *and*
- *in (1),* $\forall j \neq i$: $c_j = \frac{1}{K}$.

Switching $\underline{\alpha}$ *and* $\underline{\beta}$, *we get* $\tilde{c_i} \geq \frac{1}{K}$, *with equality holding if and only if* $c_i = \frac{1}{K}$ *and for* $j \neq i$, $c_j = K$. *Thus, if* $n \geq 3$, *then either* $\max_i \tilde{c_i} < K$ *or* $\min_i \tilde{c_i} > \frac{1}{K}$, *thus there is strict inequality after another run of the iteration.*

If we choose the maximum likelihood estimate for $\underline{\beta}$, *we can see that for an arbitrary starting* $\underline{\alpha} \in \mathbb{R}_+^n$, *the sequence* $\log \alpha$ *will remain bounded. Any accumulation point of the sequence is a fix point of the iteration, but it follows from the above that there is only one fix point, namely the maximum likelihood estimate.* □

6.3 Preferential attachment tests on model V

[N] This model was created to produce uniform distribution, therefore its S value should be invariant under edge swaps.

6.3.1 The S metric

Here are some sample results. The tests confirm our expectations, the S metric does not distinguish nascent non-directed model V graphs.

CEU eTD Collection

Table 6.1: S metric values in histogram, non-directed graph, $5|E(G)|$ edge swaps. The location of the nascent graph is signified by brackets

Model	n	κ	histogram
V	500	5	1 0 0 0 0 0 2 0 1 3 0 (5) 3 5 4 2 3 6 4 3 2 4 1 0 1 1
V	500	20	1 0 0 1 2 1 1 3 1 5 (6) 1 3 3 2 2 4 6 3 3 2 0 0 0 1
V	1500	20	2 0 2 0 3 3 3 2 3 2 2 2 4 2 3 (5) 3 3 0 2 2 1 1 0 1
V	2000	30	1 0 0 0 0 0 0 0 0 1 0 0 6 1 7 3 4 3 (10) 1 5 1 3 0 2 3

The directed case shows the same properties as the other models: swaps do decrease S, for the same reason as discussed in .

Sadly, currently the S metric is the strongest test we have for model V. Since $\forall i, j = 1, \ldots, n$, $i \neq j$

$$p_{ij} = \frac{\alpha_i \alpha_j}{1 + \alpha_i \alpha_j},$$

which is not a function of d_i or d_j, therefore the urn test is not applicable.

This concludes our results about this model.

6.4 Summary

In this chapter we introduced a model that produces uniformly distributed graphs given a degree sequence. Comparing it to the previous models, this is very different in that swaps do not alter its likelihood (if it is non-directed). Thus, if a preference statistic on a power-law degree distribution graph, say, S does not decrease due to swapping, then we reject the hypothesis that it came from a preferential model, and instead accept the alternative hypothesis that it is from the uniform distribution over the degree sequence class, and can be described by model V.

CEU eTD Collection

Chapter 7

Open questions

In the process of coming up with these results, we stumbled upon some open questions, which might be interesting subjects for further research. These are:

- Edge swap mixing times.

  P.L. Erdős, I. Miklós and L. Soukup proved in [26] that graph $G(n)$ approaches uniform distribution on its degree sequence class under the swap operation in $Poly(n)$ time, if it is bipartite semi-regular. We think the statement is expandable, quite possible to all non-directed graphs, but currently this is not proven.

- Uniform generation.

  There are many publications on how to generate graphs with a specific degree sequence with uniform distribution. To the best of our knowledge, this all use some branching process, or edge mixing (e.g. swaps). Now, given a graphic sequence $\{d_1, \ldots, d_n\}$, model V will generate all graphs with this expected degree sequence with identical probability. This implies that it might be possible to use this model for generatic graphs with a prescribed degree sequence from the uniform distribution

CEU eTD Collection

from that degree sequence class. This might involve some iterative process that can remove as well as add edges. Currently we do not see the distribution of the individual degrees generated by the model, and do not have a solution for this problem, but we find it inspiring.

- The boundaries of the convex hull.

  We have stated that the maximum likelihood equation is solvable on the interior of the convex hull identified in section 6.1. What happens on the boundary?

- Perturbation vs. edge swaps.

  For models I through IV/IVa, the graphs' likelihood changed more under urn matrix perturbations than edge swaps. This implies that given a graph, there is a limit to how much its preference structure can be altered by rewiring the edges, and suggests that there is a lower bound on how much information the degree sequence contains about the graph's connection structure. The question requires further study.

- Urn likelihood and swaps.

  Given the fact that swaps on graph G had a notably stronger effects on the value $S(G)$ when G was directed, one would naturally expect to see the same with the urn likelihood. But the results begged to differ. This is counterintuitive, and should be revisited.

- Applications.

  So-called "heavy-tail" or "fat-tail" distributions have gained much popularity in recent years, in areas as far apart as finance and biology. In some of these cases, there is an underlying network model, but far from always. History shows that graph theory was able to produce some beautiful solutions for problems that had nothing to do with graphs. It is a natural idea to investigate whether one can devise a random

CEU eTD Collection

graph mechanism for a heavy-tail distribution even in the absence of a network. Examining Zipf's law about the frequency of words in a language might be an interesting starting problem.

CEU eTD Collection

Addendum

The text for this work was already closed, when we stumbled upon an article by Chatterjee, Diaconis and Sly, "Random graphs with a given degree sequence", which in essence describes our model V (see [15] and references therein). Some of the results in that paper coincide with ours:

- Uniform distribution conditioned on a degree sequence,
- Maximum likelihood equation and its solution by iteration,
- Identifying that the solution of the maximum likelihood equation is to be found on some subset of the convex hull of all graphic sequences of a certain size.

It is pleasant to see that some of the greatest researchers in the field have come to similar conclusions. We are proud to say, however, that our results are more accurate in the finite case.

This is a good lesson that teaches one to publish new results as soon as they are ready.

CEU eTD Collection

Bibliography

[1] Aldous, D. A stochastic complex network model. *Electron. Res. Announc. Amer. Math. Soc. 9* (2003), 152–161.

[2] Barabási, A.-L., and Albert, R. Emergence of scaling in random networks. *Science 286* (1999), 509–512.

[3] Barabási, A.-L., Albert, R., and Jeong, H. Scale-free characteristics of random networks: The topology of the world-wide web. *Physica A 281* (2000), 69–77.

[4] Bickel, P., and Chen, A. A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA 106* (2009), 21068–73.

[5] Bolla, M. Noisy random graphs and their Laplacians. *Discrete Mathematics 308*, 18 (2008), 4221–4230.

[6] Bolla, M., and Tusnády, G. Spectra and optimal partitions of weighted graphs. *Discrete Mathematics 128* (1994), 1–20.

[7] Bollobás, B. The diameter of random graphs. *Trans. Amer. Math. Soc. 267* (1981), 41–52.

[8] Bollobás, B. The chromatic number of random graphs. *Combinatorica 8* (1988), 49–56.

CEU eTD Collection

[9] Bollobás, B. *Random graphs, second edition*. Cambridge University Press, 2001.

[10] Bollobás, B., and Erdős, P. Cliques in random graphs. *Math. Proc. Cambridge Phil. Soc. 80* (1976).

[11] Bollobás, B., and Riordan, O. The diameter of a scale-free graph. *Combinatorica 24* (2004), 5–34.

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

[13] Borgs, C., Chayes, J., Lovász, L., T. Sós, V., and Vesztergombi, K. Convergent sequences of dense graphs II: Multiway cuts and statistical physics. *Preprint, available at* `http://research.microsoft.com/~borgs/Papers/ConRight.pdf` (2007).

[14] Borgs, C., Chayes, J., Lovász, L., T. Sós, V., and Vesztergombi, K. Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Advances in Math. 219* (2008), 1801–1851.

[15] Chatterjee, S., Diaconis, P., and Sly, A. Random graphs with a given degree sequence. *arXiv, available at* `http://arxiv.org/PS_cache/arxiv/pdf/1005/1005.1136v3.pdf` (2010).

[16] Chung, F., and Lu, L. The average distance in a random graph with given expected degrees. *Internet Mathematics 1* (2002), 15879–15882.

[17] Chung, F., and Lu, L. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics 6*, 2 (2002), 125–145.

CEU eTD Collection

[18] Chung, F., Lu, L., and Vu, V. Eigenvalues of random power law graphs. *Annals of Combinatorics 7* (2003), 21–33.

[19] Chung, F., Lu, L., and Vu, V. The spectra of random graphs with given expected degrees. *Internet Mathematics 1*, 3 (2003), 257–275.

[20] Coja-Oghlan, A. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing 19*, 2 (2010), 227–284.

[21] Csiszár, V., Hussami, P., Komlós, J., Móri, T., Rejtő, L., and Tusnády, G. When the degree sequence is a sufficient statistic. *To appear* (2010).

[22] Csiszár, V., Rejtő, L., and Tusnády, G. Statistical inference on random structures. In *Horizons of combinatorics, Balatonalmádi, Bólyai Society Mathematical Studies, 17. Ed., E. Győri, G.O.H. Katona, L. Lovász* (2008), Springer Verlag, Berlin, pp. 37–66.

[23] Dorogovtsev, S., and Mendes, J. *Evolution of networks: from biological nets to the Internet and WWW*. Oxford University Press, 2003.

[24] Dorogovtsev, S., Mendes, J., and Samukhin, A. Structure of growing networks with preferential linking. *Phys. Rev. Lett. 85* (2000), 4633–4636.

[25] Durrett, R. *Random graph dynamics (Cambridge series in statistical and probabilistic mathematics)*. Cambridge University Press, 2007.

[26] Erdős, L. P., Miklós, I., and Soukup, L. Towards random uniform sampling of bipartite graphs with a given degree sequence. To appear.

[27] Erdős, P. On cliques in graphs. *Israel J. Math. 4* (1966), 233–234.

CEU eTD Collection

[28] Erdős, P., and Gallai, T. Gráfok előírt fokú pontokkal. *Mat. Lapok 11* (1960), 264–274.

[29] Erdős, P., and Rényi, A. On random graphs, I. *Publ. Math. Debrecen 6* (1959), 290–297.

[30] Erdős, P., and Rényi, A. On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences* (1960), pp. 17–61.

[31] Erdős, P., and Rényi, A. On the evolution of random graphs. *Bull. Inst. Internat. Statist. 38*, 4 (1961), 343–347.

[32] Frieze, A. On the independence number of random graphs. *Discrete Math. 81* (1990), 171–175.

[33] Frieze, A., and Kannan, R. Quick approximation to matrices and applications. *Combinatorica 19* (1999), 175–220.

[34] Grimmett, G., and McDiarmid, C. On colouring random graphs. *Math. Proc. Cambridge Philos. Soc. 77* (1975), 313–324.

[35] Ide, Y., Konno, N., and Obata, N. Spectral properties of the threshold network model. *arXiv, available at* `http://arxiv.org/pdf/1001.0136v1` (2010).

[36] Ispány, M., Michaletzky, G., J. Reiczigel, Tusnády, G., Tusnády, P., and Varga, K. Approximation of nonnegative integer-valued matrices with application to hungarian mortality data. In *Talk at a Budapest conference* (2010).

[37] Janson, S., Knuth, D., Łuczak, T., and Pittel, B. The birth of the giant component. *Random Structures and Algorithms 4* (1993), 233–358.

CEU eTD Collection

[38] Janson, S., Łuczak, T., and Ruciński, A. *Random graphs*. Wiley-Interscience, 2000.

[39] Keller, E. Revisiting "scale-free" networks. *Bioessays 27*, 10 (2005), 1060–68.

[40] Kohayakawa, Y. Szemerédi's regularity lemma for sparse graphs. In *Foundations of computational mathematics* (1997), Springer, pp. 216–230.

[41] Krapivsky, P., Redner, S., and Leyvraz, F. Connectivity of growing random networks. *Phys. Rev. Lett. 85* (2000), 4629–4632.

[42] Li, L., Alderson, D., Doyle, J., and Willinger, W. Towards a theory of scale-free graphs: definition, properties and implications. *Internet Mathematics 2*, 4 (2005), 431–523.

[43] Liljeros, F., Edling, C., Amaral, L., Stanley, H., and Åberg, Y. The web of human sexual contacts. *Nature 411*, 21 (2001), 908–909.

[44] Łuczak, T. The chromatic number of random graphs. *Combinatorica 11* (1991), 45–54.

[45] Matula, D. On the complete subgraphs of a random graph. In *Proc. 2nd Chapel Hill Conf. Combinatorial Math and its Applications (R. C. Bose et al., eds.), Univ. North Carolina, Chapel Hill* (1970), pp. 356–359.

[46] Matula, D. The employee party problem. *Notices Amer. Math. Soc. 19* (1972), A–382.

[47] McDiarmid, C. On the method of bounded differences. In *In Surveys of Combinatorics, Proceedings, Norwich, 1989, London Math. Soc. Lecture Note Set. 141* (1989), Cambridge Univ. Press, Cambridge, pp. 148–188.

CEU eTD Collection

[48] Moon, J., and Moser, L. On cliques in graphs. *Israel J. Math. 3* (1965), 359–368.

[49] Móri, T. F. The maximum degree of the Barabási-Albert random tree. *Comb. Probab. Comput. 14* (2005), 339–348.

[50] Nepusz, T., Négyessy, L., Tusnády, G., and Bazsó, F. Reconstructing cortical networks: case of directed graphs with high level reciprocity. In *Handbook of large-scale random networks (B. Bollobás, R. Kozma and D. Miklós)* (2009), Springer, pp. 325–368.

[51] Newman, M., Barabási, A.-L., and Watts, D. *The structure and dynamics of networks (Princeton studies in complexity).* Princeton University Press, 2007.

[52] Prüfer, H. Neuer beweis eines satzes über permutationen. *Arch. Math. Phys. 27* (1918), 742–744.

[53] Shamir, E., and Spencer, J. Sharp concentration of the chromatic number on random graphs $G_{n,p}$. *Combinatorica 7* (1987), 124–129.

[54] Szemerédi, E. On the sets of integers containing no k elements in arithmetic progression. *Polska Akademia Nauk. Instytut Matematyczny. Acta Arithmetica 27* (1975), 199–245.

[55] Szemerédi, E. Regular partitions of graphs. In *Proceedings, Colloq. Inter. CNRS (J-C. Bermond, J-C. Fourier, M. Las Vergnas and D. Sotteau, editors)* (1976), pp. 399–401.

[56] Ullman, J. An algorithm for subgraph isomorphism. *Journal of the ACM 23* (1976), 31–42.

[57] Watts, D., and Strogatz, S. Collective dynamics of 'small-world' networks. *Nature 393* (1998), 440–442.

CEU eTD Collection