Mathematical Models of the Web Graph and Integrality Ratio of the 2EC Subgraph Problem on Multigraphs

Alexander Nazarov

Dept. of Combinatorics and Optimization,

University of Waterloo, Waterloo, ON, N2L 3G1

May 17, 2005

Abstract

We present an essay composed of two parts. The first chapter surveys the mathematical models used to describe the web graph. We focus our discussion on a preferential attachment model presented by Barabasi and Albert. We discuss the Bollobas and Riordan LCD model which describes the same preferential attachment model but in strict mathematical terms. We also discuss a model presented by Aldous. An off-line model proposed by Pralat and Luczak is presented last. The second chapter investigates the integrality ratio of the 2 edge connected subgraph problem in multigraphs. We present a simpler proof of a recent theorem of Carr and Ravi. We also discuss the lower bounds on the ratio that were found through computational work on small graphs. Note that the two chapters are unrelated, but are both of interest to me.

Acknowledgements

I would like to acknowledge my supervisor Joseph Cheriyan for all of his valuable advice. I would also like to thank Pawel Pralat for our discussion about his model. Finally, I thank the official readers of my essay: Joseph Cheriyan and Ashwin Nayak, as well as the unofficial readers of my essay: Antony Bonato and Graeme Kemkes.

I would like to acknowledge the help of Mike Stilp in providing some ideas for the proof of the Carr-Ravi theorem. I wouls like to acknowledge Jordan Sehn as the initial author of the computer program that I modified to obtain the results of the integrality ratio.

Contents

Chapter 1

Introduction

The *web graph* is constructed by considering each web page as a node and each hyperlink as an edge between two nodes. It is not know how large the web graph is, but at the time of the writing it contains over 8 billion nodes [20]. Studying and understanding the behavior of the web graphs allows us to simulate it and study important properties of the graph that lead to different applications, such as optimization of search engines for better extraction of information.

It is important to note that the web graph is different from the *Internet graph*. By definition, the Internet is composed of the physical resources such as cables and routers that allow network communication to take place between computers. Each node and edge in the Internet graph is constructed by looking at the relationship of routers. The Internet graph is bounded by geographical constraints and in general may have different properties from the web graph. The web graph studies the collection of web pages on the World Wide Web (WWW) and has no such geographical restrictions. The web graph is mostly known to be a 2-level structure: internal structure of the web pages on the same host, and the structure of different hosts connecting to each other. Each web page belongs to a host and about 75% of the hyperlinks connect the pages on the same host [8]. So, the graph of web pages from the same host has a lot of structure. For example, all the pages might have a link to the main page. The *host graph* is constructed by considering all the web pages on the same host as a single vertex. In this essay, we only study the mathematical models that try to describe the

web graph, thus the entire collection of web pages.

Any mathematical model should try to have the following characteristics:

- *mathematical tractability*: one can find explicit formulas for various quantities of interest
- *fitting flexibility*: by modifying the parameters the model is able to describe the statistics observed through simulation studies on real data
- *naturalness*: the properties emerge from some simple underlying mathematical structure.

In this essay, we first discuss the studies done on the graph collected by crawling through a large collection of web pages. These empirical studies reveal interesting characteristics of the web graph. Ideally, mathematical models should be able to explain the observed properties of the graph.

We study the following three important characteristics of the web graphs which have been observed through experimental data.

- the degree distribution of nodes
- the diameter of the graph
- the clustering coefficient

The reviewed models all have different approaches to modeling the web graph. Some are done through experimental data, others through different mathematical approaches. We compare and contrast these models by pointing out their pros and cons.

In our second chapter of the essay, we discuss the problem of finding the integrality ratio for the 2 edge connected subgraph problem. There is no relation between the two topics discussed in this essay. The topic of web graph has always personally fascinated me. The topic of the 2 edge connected subgraph problem allowed me to research a very fundamental problem in combinatorial science.

Given a connected undirected complete graph $G = (V, E)$ and non-negative edge costs $c : E \to \mathbb{R}_+$, the 2EC subgraph problem is to find a multiset F of edges of minimum cost such that the subgraph $H = (V, F)$ is 2-edge connected. Note that we allow F to contain multiple copies of an edge of G. An integer programming formulation for the 2EC problem follows:

$$
\begin{array}{lllll}
\min & \sum_{e \in E} c_e x_e & & & \\
\text{s.t.} & x(\delta(S)) & \geq 2 & \forall \ \emptyset \subsetneq S \subsetneq V & \\
& x(\delta(v)) & \geq 2 & \forall \ v \in V & \text{(2EC-IP)} \\
& x_e & \geq 0 & \forall \ e \in E & \\
& x_e & \text{integral} & \forall \ e \in E &
\end{array}
$$

The LP relaxation (2EC-LP) is obtained by dropping the integrality constraints on x.

Denote the optimum objective value of (2EC-IP) by Z_{IP}, and the optimum objective value of (2EC-LP) by Z_{LP}. In general, for an integer program (IP) or a linear program (LP), we use Z_{IP} and Z_{LP} to denote the optimal values of the objective functions of the programs.

It is the goal of this essay to examine the value of the 2EC integrality ratio $\alpha = Z_{IP}/Z_{LP}$.

It is well known that the 2EC integrality ratio is ≤ 2. This result is due to Edmond's theorem on disjoint branching [17].

Carr and Ravi have claimed that 2EC integrality ratio is $\geq 6/5$, by constructing worst-case examples. They have proved that for 1/2-integral solution in the 2EC-LP, the value of α is 4/3. We present in this paper a simpler proof of this result.

We also perform computational work to find the lower bound for α on small graphs with a fixed number of vertices.

We briefly discuss another version of the 2EC problem where multi-edges are not allowed. In other words, the values of x_e are bounded by 1. Then a graph on even 4 vertices ($n = 4$) has an integrality ratio of 4/3 as seen in figure 1.1.

The value of Z_{LP} is 3/4. For Z_{IP} any feasible integral solution has $x_e = 1$

for at least 2 solid edges, thus resulting in $Z_{IP} = 1$. The value of 2EC integrality ratio is $\alpha = Z_{IP}/Z_{LP} = 1/(3/4) = 4/3$.

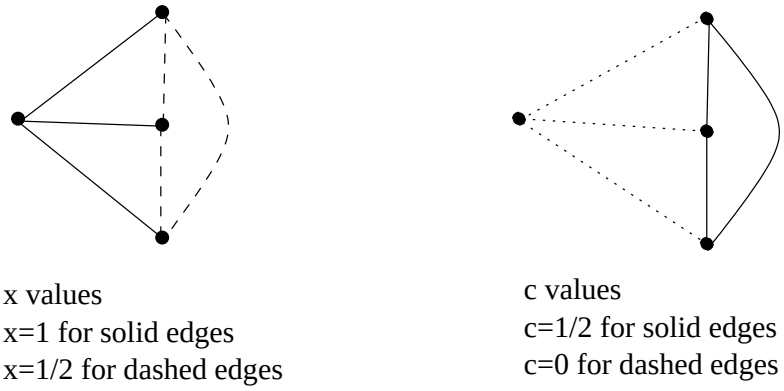

Figure 1.1: Graph on 4 vertices with $x \leq 1$ showing integrality gap of 4/3.

Now we discuss the relationship of the 2EC subgraph problem to the traveling salesman problem (TSP). Given the complete graph $K_n = (V, E)$ on n nodes with non-negative edge costs $c \in \mathbb{R}^E$, the *Traveling Salesman Problem* (TSP) is to find a Hamiltonian cycle in K_n of minimum cost. When the costs satisfy the triangle inequality constraints, we call the problem the *metric* TSP. In this paper, we only focus our discussion on the metric TSP. The integer program for the TSP is given below.

$$
\begin{array}{lllll}
\min & \sum_{e \in E} c_e x_e & & & \\
\text{s.t.} & x(\delta(S)) & \geq 2 & \forall \ \emptyset \neq S \subsetneq V & \\
& x(\delta(v)) & = 2 & \forall \ v \in V & \text{(TSP-IP)} \\
& x_e & \geq 0 & \forall \ e \in E & \\
& x_e & \text{integral} & \forall \ e \in E &
\end{array}
$$

The LP relaxation (TSP-LP) is obtained by dropping the integrality constraints on x. The new constraint $x(\delta(v)) = 2, \ \ \forall \ \ v \in V$ is called the *degree 2 constraint*.

In section 3.1 we show that the above LP is in fact equivalent to our 2EC-LP.

Part of our motivation for studying the 2EC integrality ratio comes from the following famous conjecture on TSP:

Conjecture 1.0.1 *Consider the TSP with metric costs. Then the integrality ratio of* Z_{TSP-IP}/Z_{TSP-LP} *is* $\leq 4/3$.

Moreover, the 2EC integrality ratio is less than or equal to the metric TSP integrality ratio. This holds because every integral solution of TSP-IP is also an integral solution of 2EC-IP, and the two LPs (2EC-LP and TSP-LP) are equivalent. Thus, if the above conjecture is true, then the 2EC integrality ratio is $\leq 4/3$. Note that the converse may not be true.

Chapter 2

Web Graphs

2.1 Definitions

All standard graph theory definitions are assumed as common knowledge and standard notation is used from any graph theory book. In this essay, we use the notation found in the book "Introduction to Graph Theory" by Douglas B. West [31].

We define the three characteristic properties for which we wish to compare each mathematical model.

We first define what is meant by the degree of a vertex in a graph.

Definition 2.1.1 (Degree) *The* degree of a vertex *in a graph is the number of edges incident to that vertex. The in-degree (out-degree) of a vertex is the number of edges pointing in to (out of) a vertex. We denote the values by* $d(v)$, $d_{in}(v)$ *and* $d_{out}(v)$.

Let P(k) be the probability that a randomly selected vertex has degree k. Then, P(k) follows a power law degree distribution if $P(k) \sim ck^{-\gamma}$, *where c is some positive constant, and* $\gamma > 1$ *is the power law coefficient.*

We now define two aspects of the graph that are closely related: the diameter and the average path length.

Definition 2.1.2 (Diameter) *The* diameter *of a graph G, labeled as* Diam(G) *is the maximum over all ordered pairs of vertices u and v of the shortest existing*

uv-paths. In other words, Diam(G) = $max_{u,v}$ {length of shortest uv-path | there exists a uv-path}. For an unconnected graph, the diameter is infinite.

Definition 2.1.3 (Average Path Length) *Let P be the set of all ordered (u, v) pairs such that there exists a uv-path, then the* average path length*, labeled as $< d >$, is the expected length of the shortest path, where the expectation is over uniform choices from P.*

There are two distinct definitions of *clustering coefficient* that should be mentioned.

The first definition calculates the local value of the ratio of triangles to paths of length two per each vertex v, and then takes a mean value over the whole graph.

Definition 2.1.4 (C_1) *Given a graph G we define the* clustering coefficient *as:*

$$\text{Let } C_v(G) = \frac{\text{number of triangles incident on vertex } v}{\text{number of paths of length 2 centered on vertex } v} \tag{2.1}$$

Equivalently we can also define $C_v(G)$ as follows:

$$C_v(G) = \frac{\text{number of edges between neighbours of } v}{\binom{d_G(v)}{2}} \tag{2.2}$$

Then the clustering coefficient can be defined as:

$$C_1(G) = \frac{1}{n} * \sum_{v=1}^{n} C_v(G) \tag{2.3}$$

The second definition calculates the global value of the ratio of all triangles in the graph versus the number of all pairs of adjacent vertices.

Definition 2.1.5 (C_2) *Given a graph G we define the* clustering coefficient *as:*

$$C_2(G) = \frac{3 * \text{number of triangles}}{\text{number of pairs of adjacent edges}} \tag{2.4}$$

Which can equivalently be thought of as:

$$C_2(G) = \frac{\#\text{ of pairs of } ab, ac \text{ of adjacent edges for which } bc \text{ is an edge}}{\#\text{ of pairs of } ab, ac \text{ of adjacent edges}} \tag{2.5}$$

This can be written mathematically as:

$$C_2(G) = \left(\sum_{v=1}^{n} \binom{d_G(v)}{2} C_v(G) \right) / \sum_{v=1}^{n} \binom{d_G(v)}{2} \tag{2.6}$$

Note that these two definitions are not equivalent. In particular, note the extreme example proposed by Bollobas [9]. Take G to be a double star, where vertices 1 and 2 are joined to each other and to all other vertices. None of the other vertices are connected between themselves. Then $C_v(G) = 1$ when $v \geq 3$, since for such v we have $d_G(v) = 2$, thus the denominator of equation 2.2 is $\binom{2}{2} = 1$. There is only 1 edge between the neighbours of v, so the numerator of equation 2.2 is 1, indicating that $C_{v\geq 3}(G) = 1$. Also $C_v(G) = 2/(n-1)$ when $v = 1$ or $v = 2$. To see this, we note that the degree of such v is $n-1$ so the denominator $\binom{n-1}{2} = (n-1)(n-2)/2$. There are $n-2$ edges between all the neighbours of v, thus the numerator of equation 2.2 is $n-2$. It then follows that $C_1(G) = \frac{4}{n(n-1)} + \frac{n-2}{n} = 1 - o(1)$. However, $C_2(G) \sim 2/n$.

2.2 Empirical Results

In order to try to understand what the structure of the web graph looks like, researchers have created computer programs to crawl the world wide web and construct an adjacency matrix of all the collected web pages.

Barabasi, Albert and Jeong were one of the first people who did a crawl of the web to try to discover some important properties of the web graph [6], [2]. Their study focused on collecting the web pages from the `nd.edu` domain. The size of their web graph was of about 325,000 pages and almost 1.5 million edges.

Here, we discuss in more detail the biggest publicly available study done on the web graph [13]. The study was done on an Altavista database. Altavista was one of the biggest search engines in 1999, and thus had in its index a collection of over 200 million pages and 1.5 billion hyperlinks. It is very important to note, that currently Google is the leading search engine and has over 8 billion pages in its index [20]. However, research on Google's index is proprietary information and is not published. We should also point out that crawls on the web are slightly biased to discover web pages that have a large in-degree. Web pages that no one

points to or with low in-degree to will almost never be discovered. However, the general observations about the global properties of the web graph will not be affected by this bias.

One of the key characteristics that was revealed by the above studies is that the degree distribution follows the power law with γ as the power law exponent.

Barabasi and Albert noticed that $\gamma = 2.1$ for the in-degree and $\gamma = 2.4$ for the out-degree. The study on the Altavista data indicated that $\gamma = 2.1$ for the in-degree and $\gamma = 2.7$ for the out-degree. It is worthwhile to point out that while the in-degree follows the power-law distribution very clearly, the vertices with low out-degree seem to follow a non power-law distribution.

The power-law degree distribution was a surprising result that can not be explained by classical network models such as the random graph model due to Erdos-Renyi [18, 19]. New mathematical models are needed to try to explain this important characteristic of the web graph.

In general, the web graph from the Altavista study looks almost like an organism, with the structure seen in figure 2.1.

The strongly connected component (SCC) contains about a quarter of all the web pages. The SCC is a set of nodes such that for any pair of nodes u and v in the set, there exists a uv-path. The IN and the OUT components each contain about a quarter of all the web pages. Nodes in the IN component have links that lead into the SCC component, but following the forward links from the SCC will not lead us into the IN component. The OUT component consists of pages that can be reached by following the links from the SCC, but there is no path from the OUT pages into the SCC. A real world example of nodes in the IN component is a set of newly created pages that no one has discovered but which link to other popular pages. Corporate webpages would be found in the OUT component because they only have internal links. The other quarter of the nodes are in the tendrils. Tendrils are a collection of pages that are not in the SCC, but can be reached from the IN component, or symmetrically, from which we can get to the OUT component.

An important characteristic that is revealed by the Altavista crawl is that the diameter of the strongly connected component (SCC) is logarithmic in the number of pages. The size of the SCC component is about 56 million nodes. The

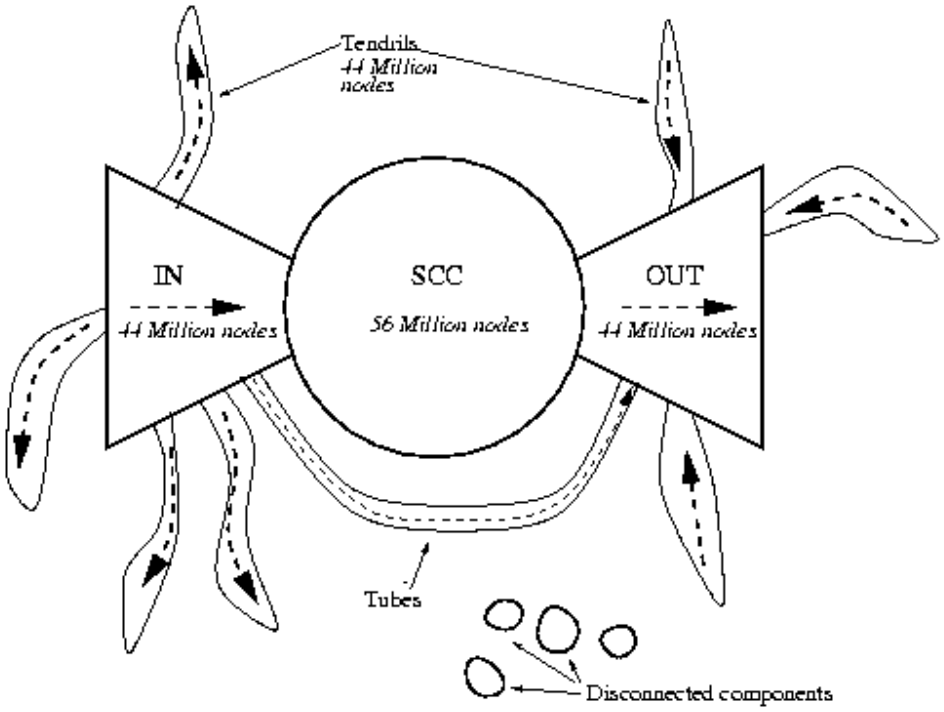

Figure 2.1: Structure of the web graph. This figure is reproduced from the Altavista study [13].

directed diameter of SCC is at least 28. However, the path from a node in the IN component to a node in the OUT component is much longer, and has a length of at least 905.

If random start and end nodes are selected, then - surprisingly - 75% of the paths do not exist in the Altavista graph. For the other 25% of the paths, the average path length is 16 for the directed version of the web graph and only 7 for the undirected version of the web graph. It is not clear if the study by Barabasi and Albert consider the `nd.edu` graph as directed or not, but they find the average path length of 11.2 on 350,000 nodes.

The small, logarithmic diameter and average path length, is another important property of the web graphs that the surveyed mathematical models aim to achieve.

The Altavista study also reveals that while 90% of the nodes belong to a large connected component, there is still a significant 10% of nodes that belong to small groups of unconnected components.

Unfortunately, the authors of the study done on the Altavista database do not calculate the clustering coefficient of the web-graph. However, the existence of clusters (communities) in the web have been noticed by other researchers. In their study, Barabasi and Albert find that $C_1(G) = 0.29$ and $C_2(G) = 0.11$.

The reviewed models try to explain the power law degree distribution, logarithmic diameter and the clustering effect that is noticed in the web graphs.

It is important to note that we do not wish for our model to describe all of the micro properties seen in figure 2.1. For example, it would be almost useless for a model to try to incorporate the notion of tendrils, since the model would have to be very complicated and would not allow for any analytical results. It is far more important for a good model to describe the overall behavior seen in the empirical studies. To make a comparison to the world of physics, the laws of Newtonian physics do not incorporate the notion of friction but describe in general the dynamics of the objects.

The analytical results of the models reviewed will almost never produce exactly the same results that are seen in the empirical studies. We perform such comparison to see which model is able to produce the best results, and also to demonstrate the ability of the model to adjust its parameters to fit the real life data.

2.3 Mathematical Models

We review quickly a classical random graph model proposed by Erdos and Renyi [18, 19]. In the model, the graph $G_{n,p}$ has a fixed number of n vertices and a probability parameter p. Each pair of vertices are connected independently with the probability p or similarly, not connected with the probability $(1 - p)$.

Most properties of the graph are solved asymptotically as $n \to \infty$. We state the results for the measures of interest to us, mainly, the degree distribution, the diameter, and the clustering coefficient.

Since each edge can occur or not occur with independent probability, the model has a Poisson degree distribution. Thus,

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{(p(n-1))^k e^{-(p(n-1))}}{k!}.$$

The expected mean degree of each vertex is $p(n-1)$.

For the diameter of the random graph, we note that $Diam(G) = \infty$ if the graph is disconnected. However, if $pn/\log n \to \infty$ and $\log n/\log(pn) \to \infty$ then with very high probability $Diam(G_{n,p}) = O(\log n/\log(pn))$

The clustering coefficient is low, since the probability of connection between two vertices is p regardless of whether they have a common neighbour. Thus, $C_1 = p$ and $C_1(G_{n,p}) \sim n^{-1}$.

Due to the fact that the random graph model does not explain the power law degree distribution that is seen in the empirical studies on web graphs, new models were developed that are reviewed in the next four sections.

The power law degree distribution was noticed much earlier in other networks. For example, Zipf's law, named after the Harvard linguistic professor George Kingsley Zipf (1902-1950) states that if all the events are ranked in increasing order or some measure of magnitude, then the probability that a random event has rank k is $P(k) \sim k^{-\alpha}$ with exponent α close to 1. This relation was noticed for the occurrence of words in a long piece of text. There was no model proposed by Zipf to try to explain this observation.

2.4 Preferential Attachment Model

The preferential attachment model was proposed by Barabasi and Albert et al. [4], [5] as a means to try to explain the power law degree distribution of the web pages. It should be noted that most of the results obtained by Barabasi et al. are through simulations. Theirs is not the first mathematical model that uses the concept of preferential attachment to explain power law degree distribution. We discuss this in more detail in section 2.4.2.

2.4.1 Motivation

Barabasi and Albert et al. were the first researchers to discover that the web graph has power law degree distribution [7]. The main motivation of this model is to try to explain the observed characteristics of the web graph, mainly the power law degree distribution and logarithmic diameter. The classical random graph model does not explain these characteristics. Thus, they tried to take a fundamentally different approach to modeling the dynamics of the network. In order to explain the empirical results, they incorporate the ideas of growth and preferential attachment into their mathematical model. The concept of growth is not present in the random graph model since the model starts with a fixed number of vertices. Surprisingly, Barabasi does not seem to acknowledge that the building blocks of their model have already been proposed by other scientists such as Price [30].

2.4.2 Model

The model has two distinctive properties: growth and preferential attachment. Growth means that we start with an initial graph G_0 at time $t = 0$, then a new vertex v_1 arrives at time $t = 1$ and attaches itself to old vertices with m edges. This process continues indefinitely. Preferential attachment defines the method in which the new vertex is connected to the vertices that are already in the graph. Quoting Barabasi and Albert [4], "To incorporate the growing character of the network, starting with a small number (m_0) of vertices, at every time step we add a new vertex with m ($\leq m_0$) edges that link the new vertex to m **different** vertices already present in the system. To incorporate preferential attachment, we assume that the probability Π that a new node will be connected to node i depends on the degree k_i [let $d(v_i) = k_i$] of node i, such that:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{2.7}$$

After t time steps, the model leads to a random network with $t + m_0$ vertices and mt edges."

We show how the model works through an example seen in figure 2.2 in the next section.

In other words, an old vertex with the highest degree has the highest chance of attaching itself to the new vertex. This is the well known phenomenon of "the rich get richer".

The only official parameter of the model is the value of m, which is the number of edges that a new vertex will attach itself to the old graph.

While the explanation of the model is very simple, it also leaves a lot of questions unanswered. There is no clear description of what the initial graph G_0 at time $t = 0$ looks like, other than the mention that the initial graph has m_0 vertices. The initial graph on m_0 vertices, must have some edges as well, otherwise, if there are 0 edges present, then the degree of each vertex is 0 and so a new vertex will have a probability of 0 of attaching itself to any old vertex in G_0. Now, suppose that G_0 has m_0 vertices and c_0 edges. Then after time t it must have $t + m_0$ vertices - which is stated correctly by Barabasi and Albert - and also $mt + c_0$ edges - which is not stated correctly.

It should be noted that this model is almost identical in description to Price's model [30]. In 1965 the physicist-turned-historian, Derek de Solla Price studied the network of citations between scientific papers. He then discovered through empirical studies that the in-degree and out-degree of the citation network followed the power law distribution. In 1976, Price proposed a general model to explain the power law distribution [30]. His idea was the probability that a newly appearing paper cites a previous paper (a new vertex attaches itself to an old vertex) is proportional to the number of citations the old paper already has (in-degree of the old vertex). He called it *cumulative advantage*. Each new vertex can be considered to start with in-degree of 1. The probability that a new edge attaches to any vertex i with degree k_i is:

$$\Pi(k_i) = \frac{k_i + 1}{\sum_j (k_j + 1)} \tag{2.8}$$

It can be seen, that this is almost identical to equation 2.7 of Barabasi and Albert.

Most of the analysis done on the Barabasi and Albert model was done through simulation and thus we first present a possible simulation of the model.

2.4.3 Simulation

In order to simulate and construct a network based on the preferential attachment model the following algorithm is executed.

```
Start with:  Some initial graph G_0
Input:  m

Repeat
     Add new vertex v_i to the graph
     Loop m times
          Attach v_i to vertex u with probability:  Π(k_u)) = k_u / Σ_j k_j
     End loop
End repeat
```

Note that due to the unclear explanation of Barabasi and Albert, we might have to ensure that a new vertex attaches itself to m different vertices.

Also note that due to the unclear explanation of Barabasi and Albert, it is not clear with what initial graph we should start. However, the simulation studies that are done for this model will still show the intended degree distribution and the diameter growth regardless of the initial graph.

The algorithm is actually very easy to implement. More importantly, we now discuss an approach that gives us linear running time with respect to the size of the network.

We show the working of the approach through an example. Suppose we have an initial graph with 4 vertices. Vertex 1 has degree 3, vertex 2 and vertex 3 have degree 2, and vertex 4 has degree 1. We then create an array with each vertex entered as many times as its degree. In our example the array looks like this: [1 1 1 2 2 3 3 4].

A new vertex 5 will be added to the graph and attached by m edges to old vertices in the network. By selecting a random element from the array as the target node we in fact select an old vertex with the needed probability for the preferential attachment to hold. So vertex 1 will be selected with probability 3/8,

vertex 2 with probability 2/8, vertex 3 with probability 2/8 and vertex 4 with probability 1/8.

Suppose our random selection chooses vertex 3 as the target. Then the new array will look like this: [1 1 1 2 2 3 3 4 5 3]. Note that the array never has to be sorted.

We show the working of the model and the simulation in the figure 2.2.

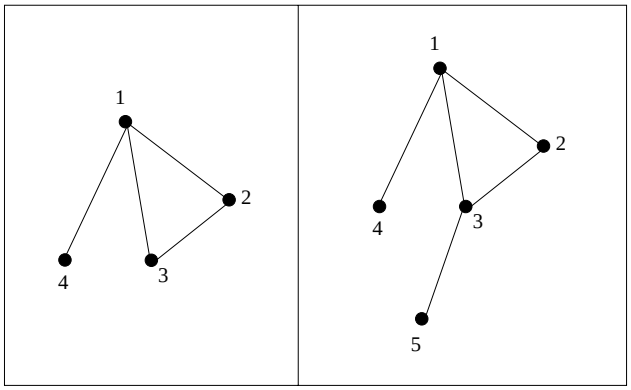

Figure 2.2: The preferential attachment model dynamics. The new vertex 5 arrives and attaches itself with a probability of 2/8 to vertex 3.

2.4.4 Analytical and Simulation Results

Barabasi and Albert perform most of the analytical results by simulating the model. While they do not explicitly state the inner working of their simulation, it most likely closely follows the outlined steps in section 2.4.3. We present here the simulation and the analytical results done on the model.

Degree distribution

Simulation results of the model show that the degree distribution follows the power law with the exponent $\gamma = 2.9 \pm 0.1$ regardless of the value of the parameter m.

Analytical results also find the power law coefficient to be $\gamma = 3$ regardless of the value of the parameter m. We quickly review how Barabasi and Albert

use continuum theory to obtain the power law coefficient. However, it has been pointed out by Bollobas [11] that the analytical results reviewed here are mathematically incorrect. We point out the inconsistencies in the calculations as we present them.

Note that it is not clear if the attachment process, as defined in the model, is repeated with the same mechanics for each of the m new edge. The description given by Barabasi and Albert seems to indicate that m edges are not added independently and all of the edges must have a different endpoint. This can be seen in the quote of their model in the previous section. However, formula 2.9 hold only if multi-edges are allowed.

All of the following calculations are quoted from [1]. Our discussion of the calculations will appear between square brackets. Assuming that $d(v_i) = k_i$ is a variable that changes according to $\Pi(k_i)$ defined in (2.7) then it follows that we wish to solve the dynamic equation:

$$\frac{\partial k_i}{\partial t} = m\Pi(k_i) = m\frac{k_i}{\sum_{j=1}^{N-1} k_j} \tag{2.9}$$

The sum in the denominator is over all the nodes except the last new node added; thus the value of the denominator is: $2mt - m$ leading to:

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t} \tag{2.10}$$

[The following mistakes need to be pointed out. First of all, the initial number of edges from the initial graph G_0 are ignored by the equation. Secondly, equation 2.9 allows for multi-edges, while the description seems to indicate that each new edge needs to be to a "different" old vertex. Finally, even if we make allowances for the first two mistakes equation 2.10 should simplify to $k_i/(2t-1)$.]

We note that the initial condition is that every node i has initial degree of m so $k_i(t_i) = m$ (degree of vertex i when it arrives at time t_i); thus the solution to equation (2.10) is:

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta} \qquad \text{with} \qquad \beta = \frac{1}{2} \tag{2.11}$$

The above equation shows us that the degree of all nodes evolves the same way, following a power law.

[It has been verified that equation 2.11 is a correct solution to equation 2.10.]

We now calculate the value of $P(k)$- the probability that a random node has degree k.

$$P[k_i(t) < k] = P\left(t_i > \frac{m^{1/\beta}t}{k^{1/\beta}}\right) \tag{2.12}$$

Assuming that we add the nodes at equal time intervals to the graph, the t_i values have a constant probability density, so:

$$P(t_i) = \frac{1}{m_0 + t} \tag{2.13}$$

Substituting this into equation (2.12) we get:

$$P\left(t_i > \frac{m^{1/\beta}t}{k^{1/\beta}}\right) = 1 - \frac{m^{1/\beta}t}{k^{1/\beta}(t + m_0)} \tag{2.14}$$

Then to get the degree distribution $P(k)$ we use:

$$P(k) = \frac{\partial P[k_i(t) < k]}{\partial k} = \frac{2m^{1/\beta}t}{m_0 + t}\frac{1}{k^{1/\beta+1}} \tag{2.15}$$

Then asymptotically ($t \to \infty$)

$$P(k) \sim 2m^{1/\beta}k^{-\gamma} \qquad \text{with} \qquad \gamma = \frac{1}{\beta} + 1 = 3 \tag{2.16}$$

Indicating that the degree distribution is independent of m.

[There are no mathematical mistakes in the above equations, if we assume that equation 2.10 is correct.]

It should be noted that the degree distribution is independent of time parameter t, and of the system size ($N = m_0 + t$). This means that despite the fact that the network keeps continuously growing it reaches a stationary power law degree distribution.

Barabasi and Albert analyze both properties of their model, the preferential attachment and the growth to see if both are needed to explain the power law degree distribution of the vertices. First, they supposed that each new vertex is attached to an old vertex with equal probability, thus keeping the growth

in the model but removing the preferential attachment property. So, $\Pi(k_i) = constant = \frac{1}{(m_0+t-1)}$.

Then continuum theory shows that the degree distribution decays exponentially following: $P(k) = \frac{e}{m}\exp(\frac{-k}{m})$. Thus, the power law characteristic disappears [1].

Without growth, the network has N nodes and no edges at the start. Each edge is added with preferential attachment. Simulations indicate that while at the beginning the model has power-law degree distribution, after N^2 steps all of the nodes become connected. The degree distribution changes from the initial power law to a Gaussian distribution.

Average path length

Barabasi and Albert simulate their model and verify the value of the average path length [2]. Computer simulations conducted by Barabasi and Albert indicate that the average path length increases approximately logarithmically with N. In particular, they find that the average path length is: $\langle d\rangle = 0.35 + 2.06\log N$ as seen in figure 2.3. For the empirical studies done by Barabasi and Albert, since the size of the network is about 300,000 nodes, then $\langle d\rangle = 11.6$, which matches closely to the observed value of of 11.2 seen in their study. Note that the simulation seems to be only performed for the parameter value of $m = 1$.

More simulations of the model performed later by Barabasi [1] show the average path length as a relation to the network size when the average degree of each vertex is equal to 4, as seen in figure 2.4. It is not clear for what parameter of m, will the average degree of a vertex be 4 in this model. They find that the average path length is: $\langle d\rangle = A\ln(N - B) + C$ for some constants A, B, C which are not specified.

Note, however, that reading off the chart for the value when $N = 10^5$ (since size of network is 350,000), we obtain an average path length of about 6, which no longer matches the empirical results seen by Barabasi and Albert.

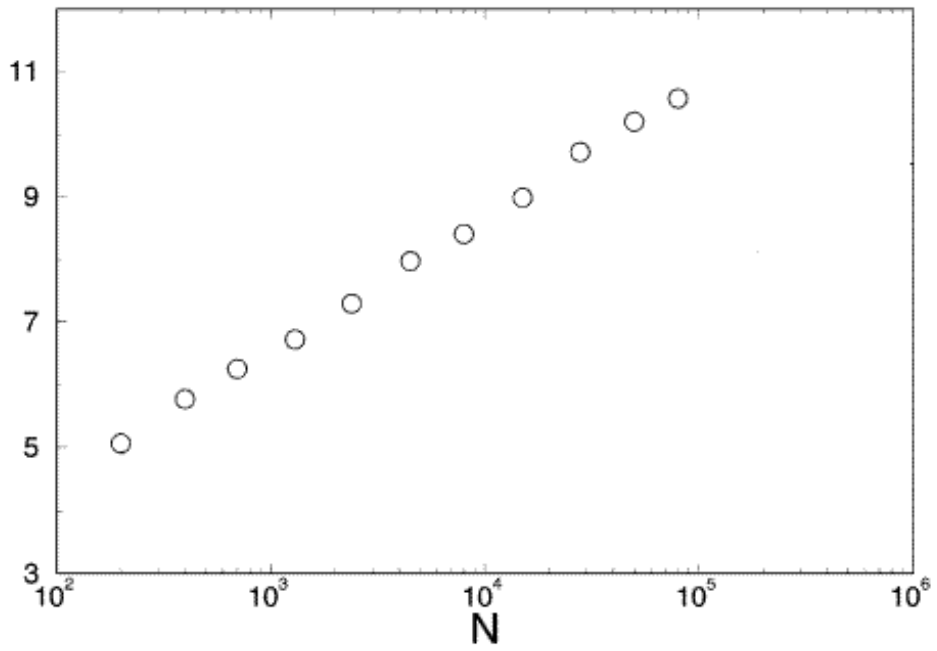

Figure 2.3: Average path length (y axis) vs. size of network (x axis) when $m = 1$. This figure is reproduced from the Barabasi, Albert and Jeong study [6] found on page 71, as figure 1.

Clustering coefficient

Barabasi and Albert perform a simulation of their model to see the relationship between the clustering coefficient and the size of the network [1]. The simulation shows that if the average degree of each vertex is kept at 4, then the preferential attachment model has a clustering coefficient that decreases approximately with power law $C = N^{-0.75}$ as seen in figure 2.5. It is not clear what definition of clustering coefficient they use.

For $N = 300,000$ this gives us a value of $C = 7.8 \times 10^{-5}$. However, the clustering coefficient in their empirical studies seem to indicate that $C_1 = 0.29$ and $C_2 = 0.11$ ([26] see Table 3.1). Therefore, it seems that the predicted model clustering coefficient does not fit the observed data.

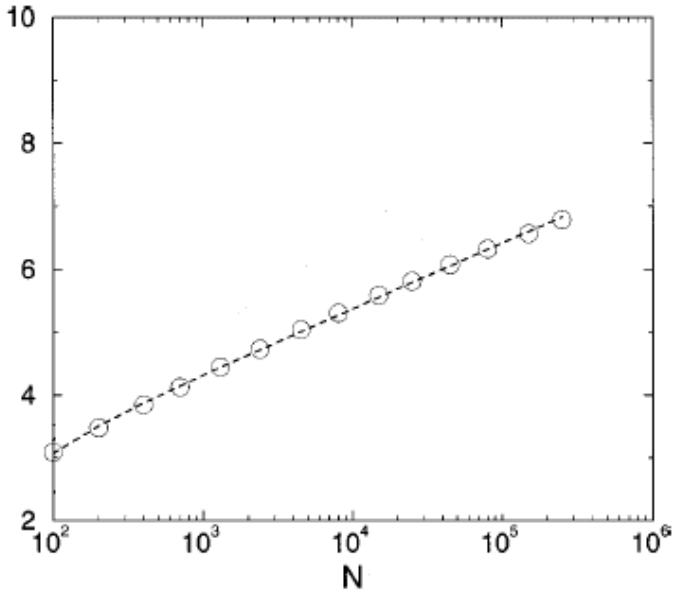

Figure 2.4: Average path length (y axis) vs. size of network (x axis) when average degree is 4. This figure is reproduced from the Barabasi and Albert study [1] found on page 74, as figure 23.

2.4.5 Pros/Cons

The preferential attachment model has a very simple and natural description incorporating growth and preferential attachment. It manages to capture some very important properties of real networks, such as the power law degree distribution and logarithmic growth. While it is not the first model to describe power law degree distribution, Barabasi and Albert were the first people to attribute the power law degree distribution to the web graph.

A drawback of the model in describing the web graph is the fact that it is undirected. While this leads to simpler mathematical results, it is not realistic. Even if we consider a slight modification to the model where a new vertex attaches itself with m directed arcs to existing vertices, then the resulting graph will be directed, but still acyclic. This modified model is in fact exactly Price's model describing the scientific citations graph [29].

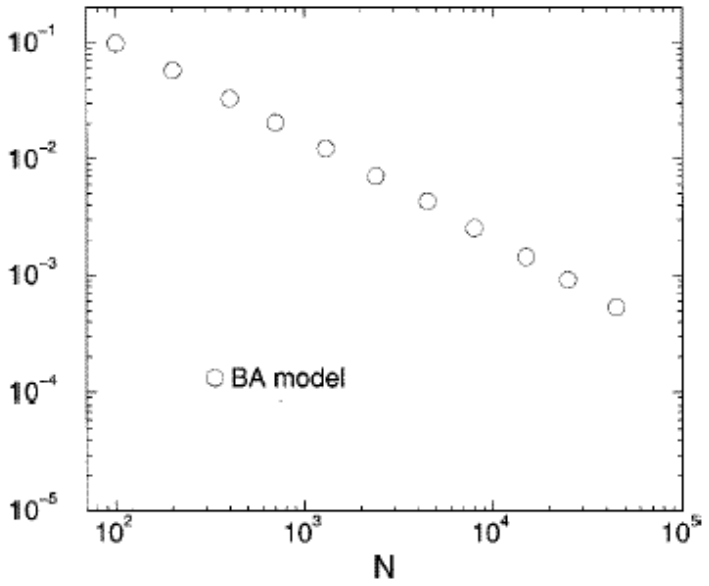

Figure 2.5: Clustering coefficient (y axis) vs. size of network (x axis) when average degree is 4. This figure is reproduced from the Barabasi and Albert study [1] found on page 75 as figure 24.

One major drawback of the original Barabasi and Albert paper, in my opinion and that of other mathematicians, is in the informal manner in which they describe their model. They are not precise about many details of their model, such as how the model behaves at the very beginning when there is only 1 vertex present. Also, it is unclear if loops and multi-edges are allowed. We examine a model by Bollobas and Riordan in the next section in which all descriptions are mathematically well defined.

The observed characteristics of the model do not fit real data. First of all, the web graph has been shown [13] to have a degree distribution with power law exponent of 2.1 (for in-degree) and 2.7 (for out-degree). The preferential attachment model only displays a power law exponent of 3 for the degree distribution. The model's only parameter of m has no effect on the degree distribution. The average path length of the model has been shown to have logarithmic growth but only for when $m = 1$. Simulations for when $m = 2$ shows that the average path

length produced by the model is less than the value that has been observed in the Barabasi and Albert empirical study. The clustering coefficient simulations for the model seem to indicate a very low clustering coefficient. Empirical studies of Barabasi et al. have shown that the real web graph has a high clustering coefficient.

Another drawback of the model is the fact that all vertices belong to one single connected component. The web graph seen in the Altavista study done by Broder et al. has one large connected component containing a large majority of the nodes and many small unconnected components.

The model does not incorporate any local events in the network. Such events can consist of node removal (a web page is no longer on the WWW), edge removal (a link from a page is removed), new edge addition after the node has been added (a link is added to the page at some later time) and a combination of any of the above (such as switching a link to point to a different page, which amounts to one edge removal and one edge addition).

The model gives a preference of attachment to older vertices. Adamic and Hubberman have argued that older web pages do not necessarily have higher degree. However, Barabasi et al. have countered, that on average, there is in fact a correlation that older web pages have higher degree [6].

2.4.6 Conclusions

The preferential attachment model introduced a new way of thinking about mathematical models to describe real networks. The model captures the concept of growth in the network and an attachment mechanism that produces the desired power law degree distribution.

The simplicity of the model resulted in its popularity. However, it leaves a lot of details unclear, and thus has incorrect (if any) analytical results. It does a poor job at describing the real dynamics of the network.

2.5 LCD Model

While the ideas of Barabasi and Albert were interesting and captured the attention of a lot of scientists, the model they propose is not well defined. We now present a model by Bollobas and Riordan [10] which is the same model as the preferential attachment model with a different mathematical point of view. The model is well defined, allows for multi-edges and loops and thus allows for rigorous analytical results to be done.

2.5.1 Motivation

Bollobas wanted to define every aspect of the preferential attachment model and present some mathematical tools that will allow exact formulas to be obtained for the degree distribution, the diameter and the clustering coefficient. In particular Bollobas and Riordan present the concept of linearized chord diagrams (LCDs) to explain the model [11].

2.5.2 Model

We discuss the model by clarifying and giving precise definitions for each detail of the preferential attachment model.

During the growth process of the model, a new vertex is added to the graph and connected by m new edges to old vertices. Just like in the preferential attachment model, m is the only parameter of the model. Multiple edges are allowed between the same pair of vertices. Loops are also allowed in the model. In the language of web pages this corresponds to a web page linking to itself (a link to the top of the same page) or a multiple links to the same page (perhaps multiple links to different sections of the same page).

In the preferential attachment model, it is not clear with which graph the model starts with. Bollobas introduces the notation G_m^t to indicate a graph G that is formed after t new vertices are added and with the model parameter m. We define $d_{G_m^t}(v)$ as the degree of vertex v of the graph G_m^t. The model starts with graph G_1^1 which consists of a single vertex with one loop.

Suppose for a moment that $m = 1$. Then the graph G_1^t is formed from the graph G_1^{t-1} by adding a new vertex v_i. A single new edge (since for now we consider $m = 1$) between the new v_t and the old v_i is added where v_i is chosen with the following probability:

$$\Pi(i) = \begin{cases} d_{G_1^{t-1}}(v_i)/(2t-1) & 1 \leq i \leq t-1 \\ 1/(2t-1) & i = t \end{cases} \tag{2.17}$$

So, an edge is added with the preferential attachment property. After t new vertices are added, we have a total degree of $2t-2$ in our graph. Since the model allows for loops the total degree in the denominator is $2t-1$.

If $m > 1$, then each new vertex v_t is connected with m edges where each neighbour of v_t is selected independently with probability defined in (2.17).

We present the linearized chord diagram (LCD) as mathematical tool that allows us to study this model. In order to explain the dynamics of the LCD model, we first concentrate on the case when the model parameter m is set to 1.

Definition 2.5.1 (LCD) *An* LCD *is an* n-pairing *that partitions the set* $\{1, 2, \ldots, 2n\}$ *into pairs, such that there are* $(2n)!/(n!2^n)$ *n-pairings.*

An LCD with n-chords has $2n$ distinct points on the x-axis paired off by semi-circular chords in the upper half plane. Each chord has a left endpoint and a right endpoint.

Seen in figure 2.6 is an example of an LCD with 6 chords and thus 12 points on the x-axis.

Given an LCD L we can form a graph $\phi(L)$ as follows. Starting from the left, vertex 1 will be formed from all endpoints up to and including the first right endpoint of some chord. In our example, the first right endpoint is 4. Vertex 2 will be formed from all endpoints up to and including the second right endpoint of some chord (in our example, vertex 2 will consist of points 5 through 8), and so on. In other words, we first identify all the right endpoints in the LCD L. Then, we shrink the interval from 0 up to the first right endpoint to form vertex 1. Then, shrink the interval from the first right endpoint to the second most right endpoint to form vertex 2, and so on. Now replace each chord of L by an edge e.

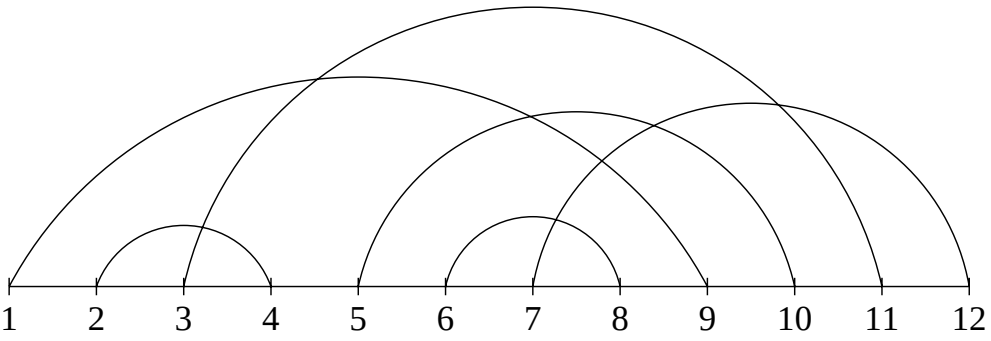

Figure 2.6: LCD with 6 chords on 12 points.

The edge e connects the vertex that corresponds to the left endpoint of a chord to the vertex that corresponds to the right endpoint of a chord.

Seen in figure 2.7 is the resulting graph $\phi(L)$=G_1^6 that we obtain from the LCD in our example. We label each vertex by the right endpoint to which it corresponds.

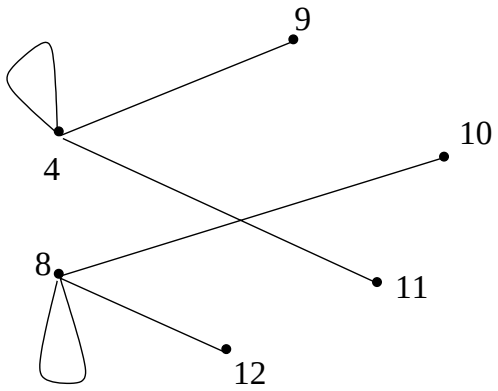

Figure 2.7: The resulting graph $\phi(L)$ on 6 vertices with 1 edge added at each step.

Suppose that L is chosen uniformly at random from all $(2n)!/(n!2^n)$ LCDs with n chords. Then to see that the graph $\phi(L)$ has the same distribution as shown in (2.17) consider the following situation. Take a random LCD $L^{'}$ with $n-1$ chords (thus $\phi(L^{'})$ has $n-1$ edges) and add a new chord whose right

endpoint is to the right of all $n-1$ chords. Place the left endpoint of the new chord into one of the $2n-1$ possible places, each chosen with equal probability. Thus, in the graph $\phi(L')$ we initially had $n-1$ vertices connected to each other by $n-1$ edges and we have added a new vertex v_n which is connected by a new edge to another vertex with probabilities related to the degrees of the old vertices, exactly as stated in (2.17).

We now discuss how the general graph G_m^t can be obtained from G_1^{tm}, for when the parameter m is set to any value [10].

Consider the LCD as a pairing of fixed random real valued points $1, 2, \ldots, 2N$ in the interval [0,1].

Let $N = nm$ and let $x_1, x_2, \ldots, x_{2N}$ be $2N$ distinct independent samples from the uniform distribution of [0,1]. We may simply pair x_{2i-1} with x_{2i} for all i. The randomness of the order in which the x_i appear in the interval between 0 and 1 guarantees that the LCD obtained from such a pairing is the uniformly random LCD. To see this note, for any set $\{x_1, x_2, \ldots, x_{2N}\}$ of $2N$ distinct elements of [0,1], all $(2N)!$ possibilities for the order in which $x_1, x_2, \ldots, x_{2N}$ take these values are equally likely.

Let $M_2(0,1)$ be a random variable corresponding to a random variable with density function $2x, 0 < x < 1$. Now, let l_i be the left endpoints of each chord and r_i be the right endpoints of each chord in our LCD. Then $\{l_i, r_i\} = \{x_{2i-1}, x_{2i}\}$ with $l_i < r_i$. But $P(x_{2i-1}, x_{2i} \leq t) = t^2$, so the r_i are independent $M_2(0,1)$ random variables. Also, given the set of rightmost endpoints $r_1, r_2, \ldots, r_N$, the random variables corresponding to the leftmost endpoints $l_1, l_2, \ldots, l_N$ are independent and each l_i is distributed uniformly on the interval $[0,r_i]$.

To form the LCD as a pairing on the set $\{1, 2, \ldots, 2N\}$ we sort the right endpoints (r_i) in increasing order and then consider the r_i between which each leftpoint (l_i) lies.

We construct the graph G_1^{mn} as follows: start with $N = mn$ independent $M_2(0,1)$ random variables, $r_1, r_2, \ldots, r_N$. Sort the right endpoints into increasing order to obtain $R_1, R_2, \ldots, R_N$, with $R_0 = 0$. Let $L_1, L_2, \ldots, L_N$ be independent, with L_i uniform on $[0, R_i]$. Then to obtain our LCD $\mathcal{L}$ we pair L_i to R_i. Since the variables $R_1, R_2, \ldots, R_N$ are already in order, then if $R_{j-1} < L_i < R_j$ then in the graph $G_1^{nm} = \phi(\mathcal{L})$ vertex i will be connected with an edge to vertex j.

i	1	2	3	4	5	6	7	8	9	10
R_i	0.1	0.4	0.5	0.55	0.6	0.7	0.75	0.8	0.85	0.99
L_i	0.05	0.09	0.41	0.11	0.42	0.12	0.56	0.3	0.32	0.76

Table 2.1: Table pairings for an LCD on N=mn where m=2, n=5 thus N=10. The values of R_i is sorted in increasing order.

To obtain the wanted graph G_m^n we have to merge vertices in G_1^{mn} into groups of m. It only matters where the m^{th}, $2m^{th}$, etc. right endpoints lie. So we are only interested in every m^{th} endpoint and the spacing between them. Let $W_i = R_{mi}$ for $1 \leq i \leq n$ and let $w_i = W_i - W_{i-1}$ with $W_0 = 0$. We also need to consider the left endpoints of each chord. Let us define a random variable $L_{i,r}$ with $1 \leq i \leq n$ and $1 \leq r \leq m$, with $L_{i,r}$ uniform on $[0, R_{(m-1)i+r}]$. We may, in fact, work only with random variables W_i and thus $L_{i,r}$ is uniform on $[0, W_i] = [0, R_{mi}]$. Thus the graph G_m^n is obtained by taking m edges from each vertex $i, 1 \leq i \leq n$, joining i to vertices $t_{i,j}, 1 \leq j \leq m$, where $t_{i,j} = k$ if $W_{k-1} < L_{i,j} < W_k$.

The following example illustrates the construction. Let $m = 2$ and $n = 5$. Then $N = mn = 10$. We select N independent $M_2(0, 1)$ random variables $R_1, R_2, \ldots, R_N$. Our L_i are selected uniformly from the range $[0, R_i]$ as seen in table 2.1.

Our LCD L pairs of each L_i with R_i and is shown in figure 2.8.

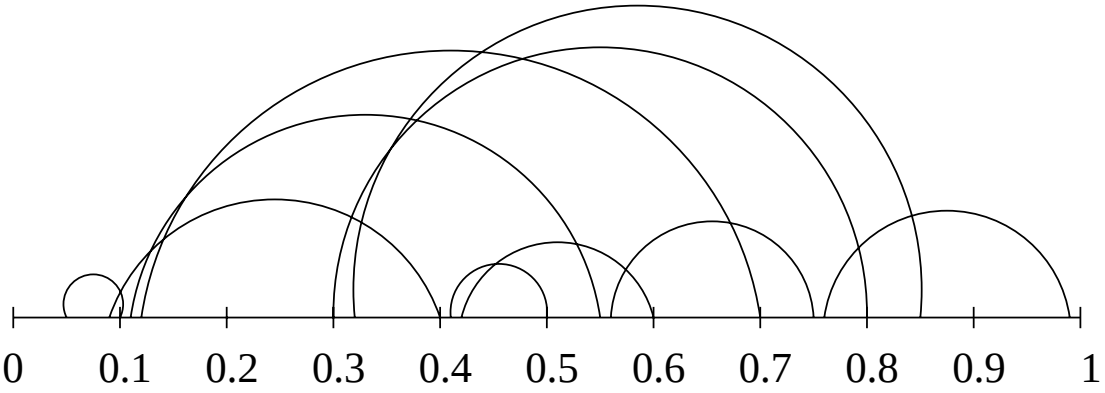

Figure 2.8: LCD on interval [0,1] with 10 arcs from Table 2.1.

i	1	2	3	4	5
W_i	0.4	0.55	0.7	0.8	0.99
w_i	0.4	0.15	0.15	0.1	0.19
$L_{i,1}$	0.05	0.41	0.42	0.56	0.32
$L_{i,2}$	0.09	0.11	0.12	0.3	0.76

Table 2.2: Table of values based on Table 2.1. Each W_i is every 2^{nd} right endpoint. Each value of w_i is the length of the interval between two right endpoints.

This gives us $G_1^{mn} = G_1^{10} = \phi(L)$ seen in figure 2.9.

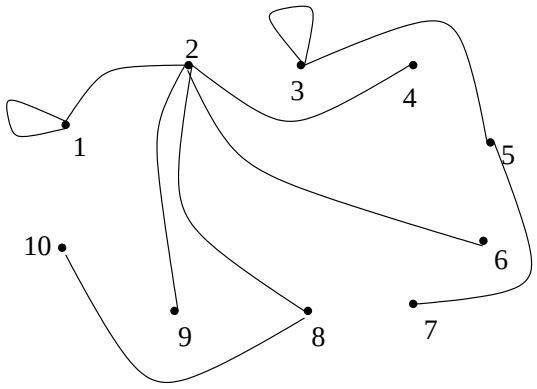

Figure 2.9: Graph on 10 vertices with 1 edge added at a time formed from the LCD in Figure 2.8.

Each W_i is every 2^{nd} right endpoint and w_i is the distance between two adjacent W_i values. The variables $L_{i,r}$ are in the range of [0, W_i]. We obtain the values seen in table 2.2.

This allows us to obtain our final G_m^n seen in figure 2.10.

We can summaries the description of G_m^n as follows: Let random variables W_i and w_i be defined as above. Given the W_i, define independent random variables $t_{i,r}$, $1 \leq i \leq n$, $1 \leq r \leq m$ with

$$P(t_{i,r} = j) = \begin{cases} w_j/W_i & j \leq i, \\ 0 & j > i. \end{cases} \tag{2.18}$$

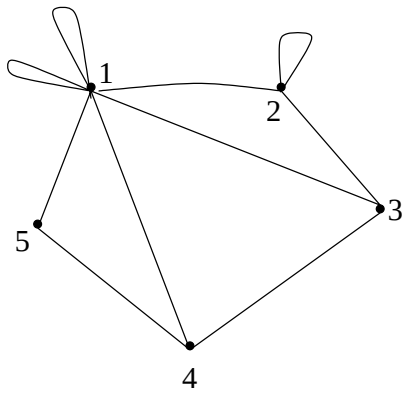

Figure 2.10: Resulting graph with 5 vertices and 2 edges added at each step obtained from Table 2.2.

Then the graph formed by taking edges from i to $t_{i,r}$ has the same distribution as G_m^n. The advantage of this approach is that we only need to be concerned about conditions on W_i.

The mathematical tools of the LCD, and pairing on the [0,1] interval allow us to perform analytical results.

2.5.3 Analytical Results

In order to analyze the graph G_m^t we can look at the snapshot picture of the graph at any time instance t. However, we mostly look at the graph when $t \to \infty$.

Given the precise specification of the model, Bollobas and Riordan derive formulas for the values of interest.

Degree Distribution

Using the description of LCDs obtained from pairings Bollobas and Riordan obtain a simple non-recursive definition of the distribution of G_m^n [11].

The following theorem describes the distribution of $P(k)$ asymptotically for all $k \leq n^{1/15}$ where n is the number of vertices in the graph.

Theorem 2.5.2 *Let $m \geq 1$ be fixed, and let G_n^m with $n \geq 0$ be the random graph*

process defined in the previous section. Let k be the total degree of a vertex. Let,

$$\alpha_k = \frac{2m(m+1)}{k(k+1)(k+2)},$$

and let $\epsilon > 0$ be fixed. Then with probability tending to 1 as $n \to \infty$ we have

$$(1-\epsilon)\alpha_k \leq P(k) \leq (1+\epsilon)\alpha_k$$

for every k in the range $0 \leq k \leq n^{1/15}$.

Note that from this theorem it is seen that $P(k) \sim 2m^2k^{-3}$ which is equivalent to equation 2.16 in the preferential attachment model.

Diameter

Calculations for the diameter were done by Bollobas and Riordan in the papers [10] and [9].

By considering an LCD as pairing on the [0,1] interval, as described in section 2.5.2. In particular Bollobas and Riordan are able to compute the formula for the diameter of G_m^n.

Theorem 2.5.3 *Let $m \geq 2$ and let $\epsilon > 0$ be fixed numbers. Then with high probability the graph G_m^n with $n > 0$ is connected and has diameter satisfying:*

$$(1-\epsilon) * (\log n / \log\log n) \leq Diam(G_m^n) \leq (1+\epsilon) * (\log n / \log\log n).$$

Note that the theorem only address the case when the model parameter is $m \geq 2$.

When $m = 1$ the graph G_m^n has a simple structure. The graph is a collection of trees, each with a loop attached. Each tree is almost identical to a structure called a *random plane-oriented recursive trees* [24]. Pittel [27] has shown that the maximum distance from any random vertex to the root of the tree is $(c+o(1))\log n$ with probability $1-o(1)$ where $c = (2\gamma)^{-1}$ such that γ is the solution of $\gamma e^{1+\gamma} = 1$. Then, we see that the diameter is $(2c + o(1))\log n$. Pittel's method in [27] can be used to prove the following theorem about G_1^n.

Theorem 2.5.4 *Let γ be the solution of $\gamma e^{1+\gamma} = 1$, and let $\epsilon > 0$ be fixed. Then for almost every G_1^n the largest distance between two vertices (diameter) in the same component is between $(\gamma^{-1} - \epsilon)\log n$ and $(\gamma^{-1} + \epsilon)\log n$.*

The Barabasi and Albert study finds that the diameter of the graph is 11.2 on 325,000 nodes. The average vertex degree is 4.51, thus $m > 1$. Then, the theorem predicts that a graph G produced by the model will have $Diam(G) = 7.44$. The Altavista graph had a SCC component of 56 million nodes and a directed diameter of 28, a directed average distance of 16 and an undirected average distance of 6.83 [13]. On 56 million nodes, the theorem predicts that a graph G produced by the model will have $Diam(G) = 8.71$.

Clustering Coefficient

Bollobas uses the LCD model to calculate the expected clustering coefficient [9]. He uses definition 2.1.5 for the clustering coefficient. Bollobas has stated and proven the following theorem:

Theorem 2.5.5 *Let $m \geq 1$ be fixed. The expected value of the clustering coefficient $C_2(G_m^n)$ satisfies:*

$$E(C_2(G_m^n)) \sim \frac{m-1}{8}\frac{(\log n)^2}{n}$$

as $n \to \infty$.

Empirical results for the clustering coefficient were only done by Barabasi and Albert et al. They find that $C_1 = 0.29$ and $C_2 = 0.11$. The average degree in the 325,000 node graph was 4.51. Bollobas assumes that the average degree of 4 happens when $m = 2$. Then, plugging in $m = 2$ and $n = 325,000$ into the above theorem we obtain an expected value of the clustering coefficient of 1.25×10^{-5}.

2.5.4 Pros and Cons

An advantage of the LCD model is that Bollobas and Riordan give very strict definitions for the model. The model allows for multiple edges and loops. By

describing the model through well defined mathematical structures such as the LCD and the random pairings on [0,1] interval, they are able to obtain formulas for the values of interest. The power law degree distribution has been shown for vertices whose degree is not too high in relation to the total number of vertices. There are explicit formulas for the diameter and the expected value of the clustering coefficient.

The disadvantages of the LCD model are mostly inherited from the preferential attachment model. The LCD model is undirected. The power law coefficient for the degree distribution is equal to 3 regardless of the parameter of m. The diameter of the network has been shown to be sublogarithmic in the model, but the Altavista study indicates that the diameter is in fact greater than $(\log n)$. The clustering coefficient formula indicates a low clustering coefficient, which does not correspond to the empirical results seen in the Barabasi and Albert study.

The model does not allow for deletion of nodes or edges.

2.5.5 Conclusion

The LCD model is very well defined and provides mathematical formulas. However, it does not provide any parameters which can be set to specific values and would allow the model to produce values of interest that are close to the real data seen through empirical studies.

2.6 Aldous Model

Aldous presents a mathematical model that tries to describe the directed version of the web graph. He bases his model on an underlying geometry and builds up the network through growth.

2.6.1 Motivation

Aldous presents a two-parameter stochastic model of random graphs that hopes to explain the web graph. The model reviewed here is a two-parameter *mean-field*

simple copying (MFSC) model. The mathematical structures which are used in the model are discussed below.

PWIT

The model uses an underlying geometry for its graph that is based on the Poisson Weighted Infinite Trees (PWIT), which we describe below.

The Poisson weighted infinite tree is defined by a construction. First of all, the PWIT is a tree. Like every tree, there is a root vertex. We label the root vertex as 0. This root vertex is given an infinite number of links to its children vertices, which is why we call it an "infinite tree". Each link has a real-valued length. The root has a finite number of vertices within any fixed distance from the root. Each link length has the law of a Poisson process $(\xi_i, 1 \leq i < \infty)$ of rate 1 on $(0, \infty)$.

Recall that a Poisson process of rate 1 on $(0, \infty)$ is defined as follows:

- The numbers of changes in non overlapping intervals are independent for all intervals
- P(some $\xi_i \in [x, x + dx]) = 1 \cdot dx$, $0 < x < \infty$. So probability of exactly one change occurring in a small interval is equal to the size of the interval.
- Probability of two events occurring in a small interval is 0.

Suppose that ξ_i are the link lengths that are sorted in increasing order such that $0 < \xi_1 < \xi_2 < \xi_3 < \ldots$. then a Poisson process of rate 1 is defined by the property

$$\xi_1, \xi_2 - \xi_1, \xi_3 - \xi_2, \ldots \text{ are independent with Exp(1) distribution}$$

Note that the exponential distribution $Exp(\mu)$ has probability density function:

$$f(x) = \mu e^{-\mu x}, \qquad 0 < x < \infty$$

So, Exp(1) has probability density function $f(x) = e^{-x},\ 0 < x < \infty$.

Now, recursively, each vertex v arising as a child of a previous vertex is given an infinite number of links to its children and these link lengths again have the law of a Poisson process of rate 1, independent of previous lengths.

A nice property of the PWIT is self-similarity: the sub-trees at each child of the root are independent copies of the PWIT itself.

An example of a PWIT is seen in figure 2.11.

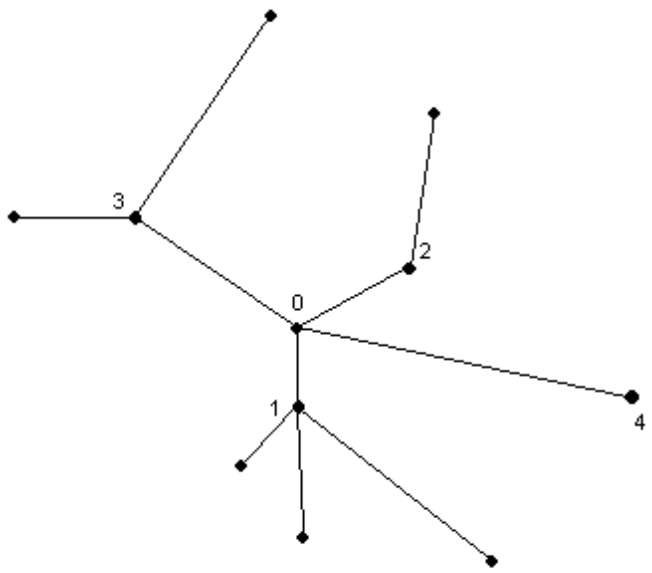

Figure 2.11: The PWIT centered at the root vertex 0, showing all children with in distance of 3 from the root. The root actually has an infinite number of children. The edge $(0, i)$ has length ξ_i. Note that $0 < \xi_1 < \xi_2 < \xi_3 < \xi_4 < 3$.

Since the PWIT is used as an underlying geometry of the model, let us explain some of the more interesting details of the structure. First of all, there is a concept of distance between two nodes in the graph. In this model, the distance has nothing to do with the number of links that have to be followed to get from one web page to another. In the language of the web pages, Aldous indicates that the distance between two web page should be thought of as the similarity of content between the two pages. For example, two pages about the topic of Golf will be close to each other.

2.6.2 Model

We first describe the dynamics of the model.

a) Vertices n = 1, 2, 3, arrive successively. Vertex n arrives at time $t(n) = \ln n$. So at some time t there are $n = e^t$ vertices. The model can be indexed by number of vertices or equivalently by time, but it is more convenient to index it by the number of vertices.

b) When vertex n arrives at time $t(n)$, there are links lengths that are defined from n to each vertex j where $1 \leq j < n$. Each link-length $d'(n, j, t(n))$ is random with exponential (mean n) law, independent of other link-lengths.

c) The link-lengths increase with time at deterministic rate 1. So at time $t > t(n)$ the link $(n,\ j)$ has length $d'(n, j, t) = d'(n, j, t(n))e^{t-t(n)}$.

d) At time t the distance $d(i, j, t)$ between two vertices $i,\ j$ (where $i, j \leq e^t =$ number of vertices present) is defined as the minimum, over all paths from i to j, of the length of the path (i.e. sum of link-lengths along the path).

Points a-d describe a geometry of an evolving random discrete metric space. Note that the geometry is the dynamic version of the PWIT. The link lengths defined by $d'(\cdot)$ follow a Poisson process of rate 1. Point (c) makes the PWIT dynamic, because the link lengths increase with time. In the language of the web graph, we can consider web pages changing constantly with time. So, as time increases, the web pages might change in content and be slowly farther away from each other in their similarity.

When a new vertex is added to the graph, it attaches itself with the following dynamics:

e) When vertex n arrives at time $t(n)$, for each $1 \leq i < n$:

   1) A directed edge (n, i) is created with probability $p(d'(n, i, t(n))$

   2) For each $j < n$ such that (i, j) is an existing edge in the graph, a directed edge (n, j) is created with probability $p(d'(n, i, t(n))$.

Two important remarks on (e).

1) The probabilities depend on link-lengths $d'(\cdot)$ and not metric distance $d(\cdot)$.

2) The probability that an existing edge (i, j) is "copied" to a new edge (n, j) depends on the link-length from n to the "tail" i, not to the "head" j.

Let us explain step (e) in the language of web pages. Step (e)(1) creates direct links from the new node to the old nodes. For example, suppose we create a new page about the topic discussed here, web graphs. We would link directly to certain sources, and with high probability we would link to the sources that are very similar to us, therefore, nodes that have small link distances from us. So, we might link to a web page of Barabasi, a web page of Bollobas, a web page of Aldous and a web page of Pralat. Step (e)(2) copies links from certain nodes. We would trust the sources of pages that are closest to us (having very similar content as us). Continuing the example, we might copy two links from a Bollobas web page. But we might also copy a link from the web page of Broder (for example to a web page about his Altavista study) even though we do not directly link to his page. This step of the model allows us to create links to pages farther away, as long as we trust the source of the link.

Note that in point (e) new edges are created with a probability function. The two parameters of the model: $\alpha, \lambda > 0$, enter via this probability function:

$$p(x) = min(1, \alpha\lambda e^{-\lambda x}), \qquad 0 \leq x < \infty.$$

It is required that

$$\int_0^\infty p(x)dx < 1. \tag{2.19}$$

Equation 2.19 holds for two parameter ranges. We show the calculations for these ranges.

Suppose $\alpha\lambda \leq 1$. Then $p(x) = \alpha\lambda e^{-\lambda x}$ for $0 \leq x < \infty$. Then,

$$\begin{aligned} lim\ n \to \infty \quad & \alpha \textstyle\int_0^n \lambda e^{-\lambda x} dx \quad = \\ lim\ n \to \infty \quad & \alpha[-e^{-\lambda x}]_0^n \quad = \\ lim\ n \to \infty \quad & \alpha[-e^{-\lambda n} + e^{-\lambda 0}] \quad = \\ & \alpha[0 + 1] \quad = \\ & \alpha \end{aligned} \tag{2.20}$$

Thus, when $\alpha\lambda \leq 1$ we require $\alpha < 1$.

Now suppose $\alpha\lambda > 1$. Then,

$$p(x) = \begin{cases} 1 & \text{if } x < \frac{log\frac{1}{\alpha\lambda}}{-\lambda} \\ \alpha\lambda e^{-\lambda x} & \text{otherwise} \end{cases}$$

Now we need to solve $\int_0^n p(x)dx < 1$.

$$\begin{aligned} lim\ n \to \infty \quad & \int_0^{\frac{\log\frac{1}{\alpha\lambda}}{-\lambda}} 1dx + \alpha \int_{\frac{\log\frac{1}{\alpha\lambda}}{-\lambda}}^{n} \lambda e^{-\lambda x} && = \\ lim\ n \to \infty \quad & \frac{\log\frac{1}{\alpha\lambda}}{-\lambda} + \alpha[-e^{-\lambda x}]_{\frac{\log\frac{1}{\alpha\lambda}}{-\lambda}}^{n} && = \\ lim\ n \to \infty \quad & \frac{\log(\alpha\lambda)}{\lambda} + \alpha[-e^{-\lambda n} + e^{-\lambda\frac{\log\frac{1}{\alpha\lambda}}{-\lambda}}] && = \\ & \frac{\log\alpha\lambda}{\lambda} + \frac{\alpha}{\alpha\lambda} && = \\ & \frac{1}{\lambda}[(\log\alpha\lambda) + 1] && (2.21) \end{aligned}$$

Thus, when $\alpha\lambda > 1$ we require $1/\lambda \cdot (\log(\alpha\lambda) + 1) < 1$.

The analytical results will always make a distinction between these two regions of solutions. Aldous calls the ranges as "low" (for $\alpha\lambda \leq 1$) and "high" (for $\alpha\lambda > 1$).

$$\begin{array}{ccc} 0 < \lambda \leq 1/\alpha & 0 < \alpha < 1 & [\text{low}] \\ \alpha\lambda > 1 & 1/\lambda \cdot (\log(\alpha\lambda) + 1) < 1 & [\text{high}] \end{array}$$

It is convenient to reparameterize the "high" region by using $\eta = \lambda^{-1}\log(\alpha\lambda)$. Thus, the parameter ranges for the high clustering region becomes:

$$0 < \eta < 1, \quad \eta + 1/\lambda < 1 \quad [\text{high}]$$

2.6.3 Analytical Results

The parameter α and λ and the value $\eta = \lambda^{-1}\log(\alpha\lambda)$ could be used to control the mean degree value and the clustering value.

Let $d_{in}(v)$ correspond to the in-degree of a random vertex and $d_{out}(v)$ correspond to the out-degree of a random vertex. Then,

$$E(d_{in}(v)) = E(d_{out}(v)) = \partial = \begin{cases} \frac{\alpha}{1-\alpha} & \text{[low]} \\ \frac{\eta+1}{1-\eta-1/\lambda}\ \lambda & \text{[high]} \end{cases} \tag{2.22}$$

This corresponds to the expected value of the average degree of a vertex.

Just like we did with η, there are other parameterizations that can be done. In particular we define a value β_u for integers $u \geq 1$ as:

$$\beta_u = \begin{cases} \alpha^u \lambda^{u-1}(1/u) & \text{[low]} \\ \eta + \frac{1}{u\lambda} & \text{[high]} \end{cases} \tag{2.23}$$

For a directed graph we define the normalized clustering coefficient κ as: the proportion of directed 2-paths $v_1 \to v_2 \to v_3$ for which $v_1 \to v_3$ is also a directed edge. The parameter κ gives an overall measure of triangle density and is a directed version of the definition 2.1.5 of C_2.

We now look at how the parameters can be used to control the clustering coefficient.

$$\kappa = \begin{cases} \frac{\alpha(1-\alpha)\lambda}{2-\alpha^2\lambda} & \text{[low]} \\ \frac{(\eta+\frac{1}{2\lambda})(1-\eta-1/\lambda)}{(\eta+1/\lambda)(1-\eta-\frac{1}{2\lambda})} & \text{[high]} \end{cases} \tag{2.24}$$

The two regions can be specified as:

$$\begin{array}{lll} 0 < \partial < \infty, & 0 < \kappa \leq \frac{1}{\partial+2} & \text{[low]} \\ 0 < \partial < \infty, & \frac{1}{\partial+2} < \kappa < 1 & \text{[high]} \end{array}$$

So the two model parameters have a direct interpretation in terms of mean degree and clustering. The regions can be thought of as "high" and "low" clustering regions.

Every pair of values of ∂ with $0 < \partial < \infty$ and κ with $0 < \kappa < 1$ occurs for a unique parameter pair (α, λ).

Let us use the Barabasi and Albert study to show an example how the values of (α, λ) or (η, λ) can be found. The mean degree is $\partial = 4.51$. The definition of κ is a directed version of C_2, thus let $\kappa = 0.29$.

First, find if we are in the "high" or "low" clustering region.

$$\frac{1}{\partial+2} = \frac{1}{4.51+2} = 0.15.$$

Since $\kappa = 0.29 > 0.15$, we are in the "high" region. Now use the mean degree to find the value of $\eta + 1/\lambda$.

$$\partial = 4.51 = \frac{\eta + 1\ \lambda}{1 - \eta - 1/\lambda} = \frac{X}{1 - X}.$$

Where we let $X = \eta + 1/\lambda$. Then,

$$X = \eta + 1/\lambda = 0.8185 \tag{2.25}$$

We now use the value of κ to find the value of η and λ.

$$\kappa = 0.29 = \frac{(\eta + \frac{1}{2\lambda})(1 - \eta - 1/\lambda)}{(\eta + 1/\lambda)(1 - \eta - \frac{1}{2\lambda})}.$$

Let $Y = \eta + \frac{1}{2\lambda}$. Then,

$$0.29 = \frac{Y}{1 - Y} \cdot \frac{1 - 0.8185}{0.8185}.$$

Then,

$$Y = 0.57 = \eta + \frac{1}{2\lambda} \tag{2.26}$$

Subtracting equation 2.26 from equation 2.25 we get $\lambda = 2.01$. Plugging this back into equation 2.25 we get $\eta = 0.331$. Thus, the unique parameter values of $(\eta,\ \lambda) = (2.01,\ 0.331)$.

Degree Distribution

We introduce some standard statistical notation for the binomial distribution ($\text{Bin}(m, p)$), geometric distribution ($\text{Geo}(p)$) and Poisson distribution ($\text{Poi}(p)$).

The geometric distribution, $\text{Geo}(p)$ is where:

$$P(i) = (1 - p)^{i-1} p \quad i = 1, 2, \ldots$$
$$E(\text{Geo}(p)) = 1/p$$

The binomial distribution, $\text{Bin}(m,\ p)$ is where:

$$P(i) = \binom{m}{i} p^i (1 - p)^{m-i} \quad i = 1, 2, \ldots, m$$
$$E(\text{Bin}(m, p)) = mp$$

The Poisson distribution, Poi(p) is where:

$$P(i) = e^{-i}p^{-i}/i! \quad i = 1, 2, \ldots$$
$$E(\text{Poi}(p)) = p$$

We have already defined the parameter β_u and we point out that when $u = 1$,

$$\beta = \begin{cases} \alpha & [\text{low}] \\ \eta + 1/\lambda & [\text{high}] \end{cases} \tag{2.27}$$

The in-degree distribution works out to:

$$P_{in}(k) \sim \beta^{-2}\Gamma(1/\beta)k^{-1-1/\beta}. \tag{2.28}$$

Where the gamma function is defined as follows:

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt. \tag{2.29}$$

So the power law coefficient is $\gamma = 1 + 1/\beta$. We continue our example from the Barabasi and Albert study. We have found in the previous section that we are in the "high" clustering region. Then $\beta = \eta + 1/\lambda = 0.8185$. Then $\gamma = 2.22$. This is pretty close to the value of 2.1 found in the Altavista study for the in-degree power law coefficient. We can set $\eta + 1/\lambda = 0.91$ to simulate the in-degree distribution with power law coefficient of 2.1.

The out-degree distribution is much more complicated to find. Aldous is not able to extract a useful explicit formula for the distribution. The out-degree of the root vertex takes on the following form.

$$d_{out}(root) = \sum_{i=1}^{\infty} \text{Bin}(1 + d_{out}^{(i)}, p(\xi_i)) \tag{2.30}$$

Where $d_{out}^{(i)}$ is the out-degrees of vertex i and acts as a random variable distributed the same as $d_{out}(root)$. Also, $d_{out}^{(i)}$ is sorted in increasing order of distance from root. The distance of vertex i from the root is ξ_i.

Aldous computes moments of the above out-degree. Suppose we fix α. In the limit $\lambda \to 0$ then $P_{out}(k) \sim \text{Poi}(\alpha)$. In the case when $\alpha\lambda = 1$ then $P_{out}(k) \sim$

Geo$(1-\alpha)$. Both of the above cases do not produce power law out-degree distribution.

In this model the distributions of $d_{in}(v)$ and $d_{out}(v)$ are independent. Also, because both $d_{in}(v)$ and $d_{out}(v)$ can take the value 0, we see that $P(d_{in} + d_{out} = 0) > 0$, which implies that the graph will typically be not connected.

Diameter

Aldous has not been able to prove any specific results about the diameter or the average path length of the graph produced by the model. It is his conjecture that the average path length and the diameter are:

$$(1 + o(1)) \cdot \frac{\log n}{\log \log n} \; as \; n \to \infty$$

He does not even prove the weaker result that the average path length is $O(\log n)$ as $n \to \infty$.

Clustering Coefficient

The value κ gives an overall measure of triangle density. Aldous uses definition of C_v (used to compute C_1 in definition 2.1.4) to compute the distribution of the clustering coefficient. A random vertex v with $d(v) = k$, will have the value of C_v of:

$$C_v(k) \sim \frac{2\beta_2}{\beta - \beta_2} \cdot \frac{1}{k} \qquad as \; k \to \infty \tag{2.31}$$

This distribution has not been computed in any empirical studies.

2.6.4 Pros and Cons

The Aldous model is the only reviewed model that generates a directed web graph. However, a disadvantage of the model is that it only produces an acyclic graph, so there are no directed cycles.

The model is based on an underlying geometry and thus, each edge has notion of a distance. The distance between two web pages has a very natural correspondence to the similarity between two web pages. In my opinion, the dynamics of the PWIT geometry are also very natural.

Aldous is able to present many analytical results about the model. I think the underlying mathematical structures are elegant and present future researchers with many tools that can be used to try to solve the open problems in this model. An advantage of the model is that analytical results can be done for all possible values of clustering (between 0 and 1) and for all possible values of the mean degree (from 0 to ∞). For comparrison, the next model reviewed (Protean model), only provides analytical results for a small subset of the parameter values.

The in-degree has been shown to have a power law distribution and the parameters allow us to set the distribution to be exactly the same as the observed empirical value.

A disadvantage of the model is that there are no explicit formulas for the out-degree distribution. There are also no formulas for the average path length or the diameter of the graph. Thus, it is not certain if these values will fit in with the simulation results.

It is highly probable that the graph generated by this model is not connected. This can be viewed as both and advantage and a disadvantage. Empirical study of Broder et al. show that real web graphs have unconnected components containing about 10% of the nodes and shows an existence of a large strongly connected component[13]. Aldous does not indicate the size of the connected component generated by the model.

The in-degree and the out-degree are independent in this model. Aldous indicates this as a disadvantage of his model.

2.6.5 Conclusion

The model and its dynamics are very natural, but are also very mathematically sophisticated. This presents a tradeoff of the model being able to explain a lot of events that occur in the real web graph, but at the same time making a lot of analytical results difficult to obtain. The model shows great promise, and future work should be done to see if the diameter and the out-degree distribution do in fact follow the empirical results.

2.7 Protean Graph model

A new random graph model called the Protean Graph Model was presented by Pawel Pralat in his PhD. dissertation [28] (written in Polish) and presented in [22] as joint work with Tomasz Luczak. The model is based on the ideas of the random graph model from Erdos-Renyi.

2.7.1 Motivation

The model is motivated by the classic random graph models but introduces parameters that allow the degree distribution to follow the power law.

There is no growth in the model, as it starts out with a fixed number of n vertices. Vertices are deleted and added back to the graph. These dynamics can be thought of as the removal of web pages from the web graph and to the addition of new web pages to the web graph. Analytical results of the model are done asymptotically as $n \to \infty$.

2.7.2 Model

We denote $P_n(d, \eta)$ as a protean graph on n vertices with set model parameters d and η. The model is controlled by two parameters, $d \in \mathbb{N}$ and η, where $0 < \eta < 1$. The parameter d is the number of edges with which a new vertex connects itself to the old vertices. The parameter η is used to allow the new vertex to have a preference of attachment to the old vertices in the graph.

We describe the dynamics of the model.

Start with any graph G with n vertices. It is even fine to start with an empty graph G with n vertices and no edges, but we will use a cycle as the initial graph in our examples. Let the vertices be labeled $1, 2, \ldots, n$. Denote $[n] = \{1, 2, \ldots, n\}$. We assign an initial permutation $\sigma : [n] \to [n]$ to the vertices. For simplicity assume that the initial permutation is simply that vertex i is assigned label i (i.e. $\sigma_0(i) = i$). We call $\sigma(i)$ the *rank* or the *age* of the vertex. Then the vertex v where $\sigma(v) = 1$ is the *oldest* and the vertex u where $\sigma(u) = n$ is the *newest*.

In each step, we randomly pick one of the vertices v to be "renewed". Thus, we delete from G all edges incident to v. This corresponds to a removal of a

random node from the network, or in the language of web graphs, a removal of a web page. Suppose the selected vertex v had rank k, so $\sigma_{j-1}(v) = k$. The removed vertex gets moved to the end of the permutation, and thus becomes the *newest* vertex, so $\sigma_j(v) = n$. All other vertices who had rank higher then k move up in the permutation by one, so $\sigma_j(u) = \sigma_{j-1}(u) - 1$ for all u where $\sigma_{j-1}(u) > k$.

Now we generate d new edges - one by one - incident to the new vertex v. This can be viewed as a new node that is inserted into the graph and which establishes connections with some nodes in the network, or in the language of the web graphs, as a new web page creating link to existing web pages. In each of these d independent choices, each neighbor of v (vertex w) is chosen with probability proportional to the rank of w raised to the power $-\eta$. Authors state that it seems natural to assume that old web pages of small rank are more attractive to newcomers.

The figure 2.12 demonstrates the dynamics of the model through an example.

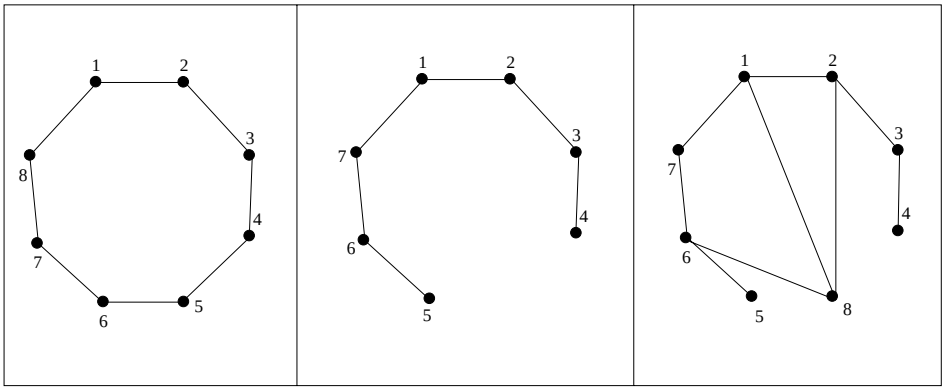

Figure 2.12: Initial graph G with vertex 5 chosen to exit and re-enter. The vertex has a new rank of 8 when it reenters the graph.

If each vertex is renewed at least once, the random graphs appearing during the process are identical random objects whose properties do not depend on the graph G we start with. It takes a bit of time for the process to reach a stationary distribution. By the "coupon collector problem" we see that after an expected number of $O(n \ \ln n)$ steps each vertex is renewed at least once.

To see this, we state the Coupon Collector Problem as follows: Given N different objects, how many tries will it take to select each one of the objects at least once if we choose each object with equal probability and with replacement. Then the expected time to select each object at least once is:

$$E[\textit{time to select } N \textit{ different objects}] = 1 + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + N \quad (2.32)$$

To see this consider the following. On our first selecting we only need 1 try to select a new object that has never been seen before. On our second selection, there are N objects to select from, and only $N-1$ objects that have never been seen, so that gives an expected number of $\frac{N}{N-1}$ trials before we select some new object. At the very end, we will only have 1 object that has never been selected, and N objects to select from, therefore it will take an expected N trials before we find the object we are looking for.

Equation 2.32 can be simplified to:

$$\begin{aligned}
N \sum_{i=0}^{N-1} \frac{1}{N-i} &= \\
N \sum_{i=1}^{N} 1/i &= \\
N \int_{1/N}^{1} \lfloor xN \rfloor^{-1} dx + 1/N &= \\
N\ (1+o(1)) \int_{1/N}^{1} \frac{1}{xN} dx + 1/N &= \\
N\ (1+o(1)) \int_{1/N}^{1} \frac{dx}{x} + 1/N &= \\
N\ (1+o(1))\ (ln\ 1 - \ln 1/N) + 1/N &= \\
N\ (1+o(1))\ (\ln N) + 1/N &= \\
(1+o(1))\ (N\ \ln N) &
\end{aligned} \quad (2.33)$$

Thus, the expected time to select N different objects is $O(N\ \ln N)$.

2.7.3 Analytical Results

The model has a fixed number of nodes in the graph. This static property of the protean graph model allows for easier analytical results to be done.

Degree Distribution

We first state the exact theorem as proven in [22] about degree distribution.

Theorem 2.7.1 *Let $\eta \in (0,1)$, $k \geq \log^2 n$, and $d = o(k)$. Then with probability tending to 1 as $n \to \infty$*

$$P(d(v) \geq k) = (1 + o(1)) \left(\frac{1-\eta}{1+\eta} \cdot \frac{d}{k} \right)^{1/\eta}.$$

Note that the theorem only states the formula for degree distribution for vertices whose degree is greater than k. It can be shown that $P(k) \sim k^{1-1/\eta}$.

For the web graph the in-degree k has power law tail with $k^{-2.1}$. The out-degree k has power law tail with $k^{-2.7}$. So, $\gamma = 2.1 = 1 + 1/\eta$ means that we can set η to 0.91 if we want the model to simulate the in-degree distribution. Similarly, $\gamma = 2.7 = 1 + 1/\eta$ means that we can set η to 0.588 if we want the model to simulate the out-degree distribution.

Diameter

We state the exact theorem as proven in [23] about the diameter of the graph produced by the protean graph model.

Theorem 2.7.2 *Let $d \geq 13$, and $0.58 \leq \eta \leq 0.92$. A protean graph has one giant component, containing a positive fraction of at least $n^{2/3}$ vertices, whose diameter is equal to $\Theta(\log n)$. The remaining components have $O(\log n)$ vertices.*

Note that η falls in the parameter range for the data to fit the degree distribution seen in the web graphs. The value of the parameter $d \geq 13$ is limiting. The average degree of a vertex is 10.46 in the Altavista graph, and thus there are many vertices who attach themselves with less then 13 new vertices. It is

possible that the mathematical results could be shown to hold for values of d much smaller.

It is interesting to note that this models allows for a single giant connected component as well as for smaller components not connected to the giant component. The Altavista graph indicates that indeed there is one large connected component as well as small unconnected components that exist.

Clustering Coefficient

The clustering coefficient of the protean graphs is not studied by the authors. However, through personal communications it seems that the model has no explicit clustering preference. Simulation results show that the clustering coefficient is low, and thus will not correspond to a realistic value found in web graphs.

2.7.4 Pros and Cons

A disadvantage of the model is that it is for an undirected graph.

The advantage of the protean graph model is that it tries to incorporate the concept of node deletion and node insertion in its model dynamics. However, a disadvantage is that the deletion and insertion is limited, and the number of nodes in the graph does not grow or shrink as time goes to infinity.

The analytical results on the model provide formulas for the degree distribution and the diameter of the graph. The parameter η can be used to set the degree distribution to be the same as observed through empirical studies. The graph has a large connected component and some unconnected components. This fits in with the empirical results from the Altavista graph. The diameter of the graph is logarithmic in growth, which is realistic. The results on diameter calculations are limited to when the parameter of $d \geq 13$, which may not hold in real networks.

Analytical results are not done on the clustering coefficient.

The model adds edges with priority to vertices with low rank (older vertices). We have discussed this assumption in the preferential attachment pros and cons, in section 2.4.5.

2.7.5 Conclusion

The model has simple well defined dynamics. It tries to incorporate the concept of deletion and addition of web page in the web graph. However, the proposed model has no growth. The degree distribution follows a power law and a parameter of the model allows us to control the power law coefficient. The analytical results are a bit limited in their range of parameter values for the diameter and degree distribution.

2.8 Concluding Remarks

The world wide web is an important technological innovation that has become a part of almost everyone's life. Yet, the structure and even the properties of the web graph remain largely unknown. Mathematical models are created to try to explain this important phenomenon. Any mathematical model should try to provide explicit formulas for various quantities of interest. The model should also provide parameters that can be used to fit real data. Finally, a model should have a natural description explaining the dynamics of the web.

We summarize the reviewed models in the table 2.3.

	BA	LCD	Aldous	Protean
Directed	No	No	Yes	No
Growth	Yes	Yes	Yes	No
Parameters	1	1	2	2
Power law Degree Distribution	Yes ($\gamma = 3$)	Yes ($\gamma = 3$)	Yes (in); ?? (out)	Yes
Log Diameter	Yes	Yes	??	Yes
High Clustering	No	No	Yes	??

Table 2.3: Summary of all the models.

Out of the models reviewed here, none of the models have all the required properties. The simpler models provide a rich variety of analytical results, but do not fit in with the empirical data. The more sophisticated models explain

some of the real web graph dynamics better, but are more difficult to understand and perform analytical results on.

There is currently no model that accepted by everyone as the authoritative source which fully describes the web graph. It would be beneficiary to see more large publicly available studies being done on the web graph. This would allow us to see the progression of some of the properties of the web graph over time, and provide better data with which to analyze and create mathematical models.

Chapter 3

Integrality ratio of the 2EC problem on multigraphs

3.1 Carr-Ravi Result

Given a connected undirected complete graph $G = (V, E)$ and non-negative edge costs $c : E \to \mathbb{R}_+$, the 2EC subgraph problem is to find a multiset F of edges of minimum cost such that the subgraph $H = (V, F)$ is 2-edge connected. Note that we allow F to contain multiple copies of an edge of G. An integer programming formulation for the 2EC problem follows:

$$\begin{array}{lllll} \min & \sum_{e \in E} c_e x_e & & & \\ \text{s.t.} & x(\delta(S)) & \geq 2 & \forall \ \emptyset \subsetneq S \subsetneq V & \text{(2EC-IP)} \\ & x_e & \geq 0 & \forall \ e \in E & \\ & x_e & \text{integral} & \forall \ e \in E & \end{array}$$

The LP relaxation (2EC-LP) is obtained by dropping the integrality constraints on x. Denote the optimum objective value of (2EC-IP) by Z_{IP}, and the optimum objective value of (2EC-LP) by Z_{LP}. Suppose that the optimum values of x^* in (2EC-LP*) only have values 0, 1/2 or 1. Denote the objective value obtained by this half-integral optimal solution as $Z_{LP(1/2)}$. Denote the edge incidence vector for a given subgraph H by χ^H.

First we show that if the cost vectors do not follow the triangle inequality relationship, then we can replace the graph by its *metric completion* (where c_{ij} for each edge takes on the value of the shortest cost $ij - path$). Note that if an edge that was made cheaper in the metric completion appears in the optimal solution of (2EC-IP), we can replace it by the shortest cost path connecting node i and j without increasing the objective value. Thus, WLOG, we can assume that the cost vectors follows the triangle inequality constraints.

We add a constraint $x(\delta(S)) = 2 \ \ \forall v \in V$ to obtain the LP (2EC-LP*).

$$\begin{array}{lllll} \min & \sum_{e \in E} c_e x_e & & & \\ \text{s.t.} & x(\delta(S)) & \geq 2 & \forall & \emptyset \subsetneq S \subsetneq V \\ & x(\delta(v)) & = 2 & \forall & v \in V \\ & x_e & \geq 0 & \forall & e \in E \end{array} \qquad \text{(2EC-LP*)}$$

We now show that if the costs follow the triangle inequality relationship (that is if $c_{ij} \leq c_{ik} + c_{jk}$ for all distinct $i, j, k \in V$) then there is an optimal solution to (2EC-LP) that is also feasible and thus optimal for (2EC-LP*).

Splitting-Off Theorem: Given a connected, Eulerian graph G and any vertex v, there exists an edge pair vx, vu such that $G' = G - vx - vu + xu$ has $| \ \delta_{G'}(S)| = | \ \delta_G(S)|$ for every node set $S \subseteq (V - \{v\})$. Moreover, if G is k-edge connected and $| \ \delta_G(v)| \geq k + 2$, then G' is k-edge connected.

Theorem 3.1.1 *In a graph $G = (V, E)$ with costs following the triangle inequality constraints, the optimum objective value of (2EC-LP*) is the same as the optimum objective value of (2EC-LP).*

Proof. Let x^* be the optimum solution of (2EC-LP). If for all $v \in V$ the constraint $x^*(\delta(v)) = 2$ holds, then we are done. Suppose there are vertices for which this constraint does not hold. Multiply x^* by some large value M such that each entry

entry in the $x*$ vector is an even integer. Let H by the multi-graph of Mx^*. Let v' be any node where $x^*(\delta(v')) \geq 2$ in G and thus degree of v' in H is $\geq 2M + 2$.

Since the degree of each vertex is even, then H is a Eulerian graph. We apply the Splitting-Off Theorem to split off the edges incident to v' to lower its degree by 2. The splitting off operation preserves the edge connectivity of the graph H. Also, the splitting off operation will replace some edges uv', yv' (where u and y are some two neighbours of v') with a single edge uy. Since the costs follow the triangle inequality, the overall cost of the objective function does not increase by the operation. We repeat the splitting off operation until $x^*(\delta(v')) = 2M$ for all $v' \in H$. Now scale back by dividing the x^* vector by M to get x satisfying (2EC-LP*). ■

We now state the main theorem that we will prove. This result is due to Carr and Ravi [14]. Our proof is a simplification of their proof and relies on the splitting-off theorem, which is not used explicitly by the Carr-Ravi proof. Moreover, our proof uses a weaker result (in 3.1.3) than theirs, because we allow multiple copies of an edge to be used in our proof.

Theorem 3.1.2 *The optimum value of 2EC-IP is within 4/3 of the half-integral solution of the 2EC-LP. So,* $Z_{IP} \leq 4/3 * Z_{LP(1/2)}$

In order to prove this theorem, we prove another result.

Theorem 3.1.3 *Let* $G = (V, E)$ *be a 4-edge connected, 4-regular graph. Let* e *by any fixed edge of* G*. Then*

$$2/3 * \chi^{G-e} = \sum_i \lambda_i \chi^{H_i}. \tag{3.1}$$

Where each H_i *is a 2EC subgraph where we allow multiple copies of the same edge to appear. Also,* $\lambda_i \geq 0$ *and* $\sum_i \lambda_i = 1$*.*

Theorem 3.1.3 implies Theorem 3.1.2. To see this, let x^* be a 1/2-integer solution for (2EC-LP*). Construct a multigraph $G(x^*)$ where we take two copies of each edge for which $x^*_e = 1$. We can think of this as multiplying the LHS and the RHS of (2EC-LP*) by 2. Thus, G is a 4-regular, 4EC multigraph.

Let us think of $\sum_i \lambda_i \chi^{H_i}$ as a weighted average over all 2EC subgraphs. Then the cost of the cheapest 2EC subgraph will satisfy: $2/3 * \chi^{(G-e)} * (cost) \geq$ cost of cheapest 2EC. But $2/3 * \chi^{(G)} * (cost) = 4/3 * Z_{LP(1/2)}$ by our construction of G. So $2/3 * \chi^{(G-e)} * (cost) = 4/3 * Z_{LP(1/2)} - 4/3 * c_e$ giving us a stronger statement that in Theorem 3.1.2.

We now prove Theorem 3.1.3 by induction.

Proof. $G = (V, E)$ is a 4-edge connected, 4-regular graph. We prove the theorem by induction on the number of vertices in G.

Base Case: $|V| = 2$.

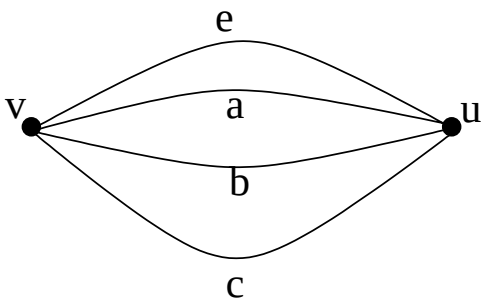

Figure 3.1: A 4-regular 4EC graph on 2 vertices.

Let e by one edge uv. Let a, b, c by the other 3 edges as seen in figure 3.1. Now we show $2/3 * \chi^{G-e} = \sum_i \lambda_i \chi^{H_i}$.

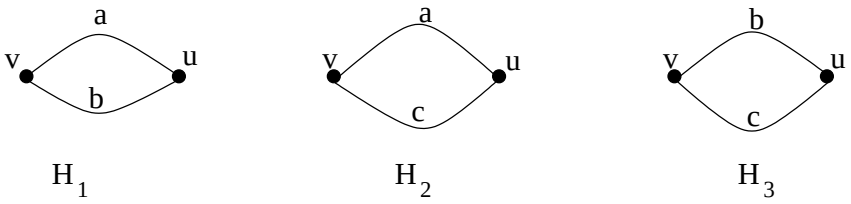

Figure 3.2: Decomposition into three 2EC subgraphs.

Then H_1, H_2, H_3 are all 2EC subgraphs of G as seen in figure 3.2. Let $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$. Then $1/3\chi^{H_1} + 1/3\chi^{H_2} + 1/3\chi^{H_3}$ implies that $x_{ab} =$

$x_{ab} = x_{bc} = 2/3$. Therefore, $2/3 * \chi^{G-e} = \sum_i \lambda_i \chi^{H_i}$ and our base case holds.

Inductive Hypothesis: Suppose the statement hold true for $|V| = k$.

Inductive Step: Show the statement to be true for $|V| = k + 1$.

Let $e = uv$ be the edge specified in the Theorem statement. Since G is 4-regular, then v has 4 neighbours which may not all be distinct.

The configurations seen in figure 3.3 can occur.

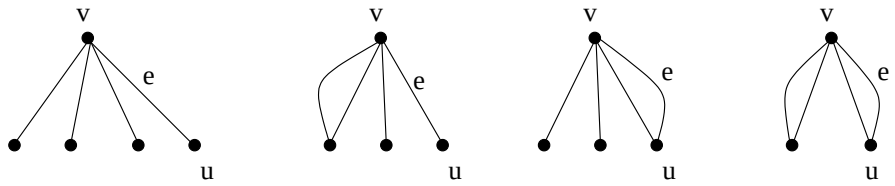

Figure 3.3: Unlabeled 4 distinct cases that can occur.

In particular, note that the there can not be 3 edges between vertices v and u, since the cut $\delta(\{uv\})$ is a cut of size 2 as seen in figure 3.4. This can not occur, since G is 4EC and thus any cut should be of size at least 4.

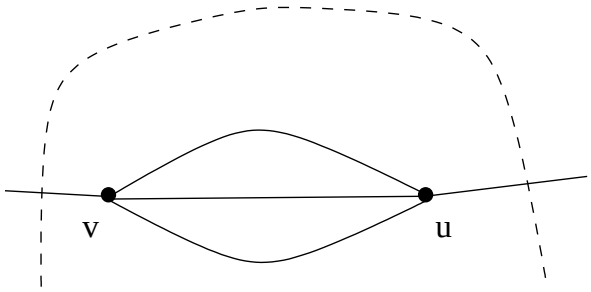

Figure 3.4: Cut of size 2 if there exits 3 edges between two vertices.

We apply the Splitting Off Theorem twice to split off all the neighbours of v. Label the neighbours of v as u, x, y, z. Without loss of generality we label the two new edges obtained through the splitting off operation as ux and yz. If u is

not distinct then we let $u = z$ and similarly, if x is not a distinct neighbour of v then let $x = y$.

So we have the situation seen in figure 3.5.

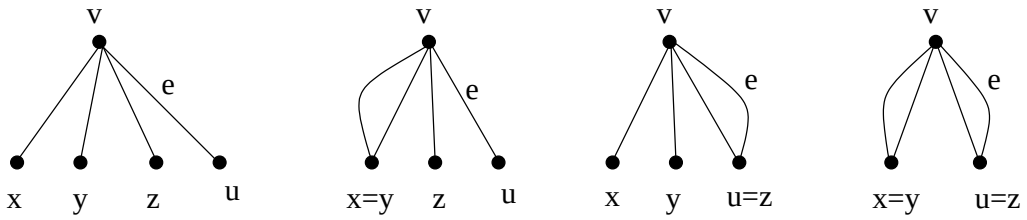

Figure 3.5: Labeled 4 cases that can occur.

Then $G_1 = G - v + ux + yz$ is a 4-regular, 4EC graph. Let $e_1 = ux$ in G_1. G_1 has one less vertex than G thus by the inductive hypothesis:

$$2/3 * \chi^{G_1 - e_1} = \sum_i \lambda_i \chi^{H_i}. \tag{3.2}$$

Define

$$\begin{aligned} H_i^1 &= H_i - yz + yv + zv \quad \text{for } yz \in H_i \\ H_i^2 &= H_i + xv + xv \qquad\quad \text{for } yz \notin H_i \end{aligned}$$

Where 2 copies of the same edge $\{xv\}$ are taken in H_i^2.

Each edge has weight of $2/3$ on the LHS of equation 3.2 thus on the RHS of equation 3.2 we have $\sum \lambda_i = 2/3$ over subgraphs H_i that contain the edge $\{yz\}$. Label such sum as λ_{yz}. Then $1 - \lambda_{yz} = 1/3$.

We claim that:

$$2/3 * \chi^{G-e} = \sum_i \lambda_i \chi^{H_i^1} + \sum_i \mu_i \chi^{H_i^2} \tag{3.3}$$

To see this, note that the only edges that appear in G that do not appear in G_1 are $\{yv\}, \{zv\}$ and $\{xv\}$. So, for all the other edges the equation 3.3 holds by the inductive hypothesis. Now, since $\lambda_{yz} = 2/3$ then $\sum_i \lambda_i$ over subgraphs H_i^1 is also equal to $2/3$. Thus the total weight of edges $\{yv\}, \{zv\}$ is $2/3$. Similarly, since $1 - \lambda_{yz} = 1/3$, then $\sum_i \mu_i$ over subgraphs H_i^2 is $1/3$. We take two copies of the edge $\{xv\}$ for all such H_i^2 and thus the total weight of edge $\{xv\}$ is $2/3$.

The result is show to hold by the principles of mathematical induction. ■

3.2 Lower Bound

We wish to find the exact integrality gap for 2EC problem on small graphs with a fixed number of vertices n.

We work with a complete graph $K_n = (V, E)$.

Note that for $n \leq 5$ the ratio is $\alpha = 1$, since the integrality ratio for metric TSP is know to be 1 for $n \leq 5$ [12].

We wish to solve the following quadratic program. Not that we have two variables per edge; informally c is a cost vector that gives the worst-case integral ratio, and x is the optimum solution to the 2EC-LP for that c.

$$
\begin{array}{lllll}
\min & \sum_{e \in E} c_e x_e & & & \\
\text{s.t.} & x(\delta(S)) & \geq 2 & \forall \ \emptyset \subsetneq S \subsetneq V & (1) \\
& x(\delta(v)) & = 2 & \forall \ v \in V & (2) \\
& x_e & \geq 0 & \forall \ e \in E & (3) \\
& c(H) & \geq 1 & \forall \text{ 2EC subgraphs H} & (4) \\
& c_{ik} + c_{kj} & \geq c_{ij} & \forall i, j, k \in V & (5) \\
& c_e & \geq 0 & \forall e \in E & (6)
\end{array}
\qquad \text{(QP)}
$$

Note that there are an exponential number of constraints of type (1) and (4)

For a particular 2EC-IP we can divide all the edge costs by the optimum value Z_{IP} and the new costs will still satisfy all the cost constraints, which are constraints (4), (5) and (6). The new costs result in a new value of $Z_{IP} = 1$. Thus, to find the integrality ratio it is sufficient to only consider metric cost function c for which the $Z_{IP} = 1$ so $\alpha = 1/Z_{LP}$.

To solve the QP-LP we do the following. Let $X = \{x_{(1)}, x_{(2)}, \ldots, x_{(p)}\}$ be the complete list of all the vertices of the 2EC-LP polytope. From polyhedral theory, for every cost function $c : E \to \mathbb{R}_+$ there exists at least one vertex $x^* \in X$ such that c is minimized over 2EC-LP polytope at x^*. (i.e such that $Z_{IP} = cx^*$).

We can now break up our set of cost functions into different not disjoint sets. For each vector $x_{(i)}$ we solve:

$$
\begin{array}{lllll}
Z_{IP_i} = \min & \sum_{e \in E} c x_{(i)} & & & \\
\text{s.t.} & c(H) & \geq 1 & \forall \text{ 2EC subgraphs H} & (C_i\text{-LP}) \\
& c_{ik} + c_{kj} & \geq c_{ij} & \forall i, j, k \in V & \\
& c_e & \geq 0 & \forall e \in E &
\end{array}
$$

Then 2EC integrality ratio is $\alpha = 1/min_i\{Z_{LP_i}\}$.

To implement this strategy as a computer algorithm we do the following steps.

Step 1: Generate the set X - all vertices in the 2EC-LP polytope. We do this using PORTA [15]. While PORTA is able to produce all the points in the 2EC-LP polytope for $n = 6$, it fails to do so for values of $n \geq 7$.

Step 2: Reduce the set of vertices X by removing all isomorphic support graphs (where $x_e > 0$). We do this using a package called nauty [25]. Modifications have been made to allow nauty to work with weighted graphs, even though it does not do this by default.

Step 3: Solve the systems C_i-LP to get a valid cost vector. We do this using CPLEX.

Step 4: Find the minimum value of Z_{LP_i} from step 3 to get our desired ratio α.

The above method was used to obtain a ratio of 10/9 for graph seen in figure 3.6 on 6 vertices.

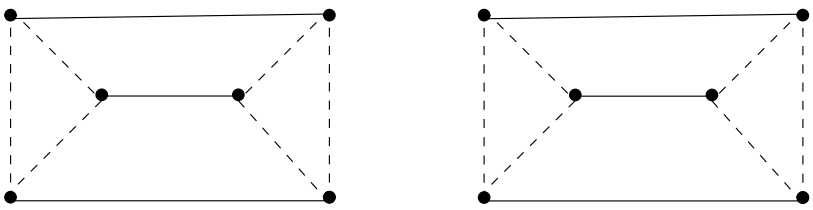

Figure 3.6: Example on 6 vertices.

Moreover, using some computer results from Boyd and Labonte [12] we got an integrality ratio of 7/6 for the graph on 9 vertices seen in figure 3.7.

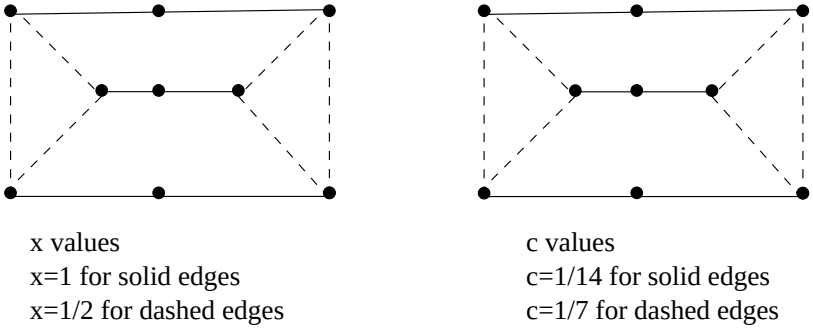

Figure 3.7: Example of 9 vertices.

Further works needs to be done to see if the same pattern of graphs will lead to a lower ratio.

Bibliography

[1] R. Albert, A.-L. Barabasi, Statistical mechanics of complex networks, *Review of Modern Physics,* **74** (2002) pp.47-97.

[2] R. Albert, H. Jeong, A.-L. Barabasi, Diameter of the world wide web *Nature,* **401** (1999) pp.130-131.

[3] A.-L. Barabasi, *Linked: The New Science of Networks*, Perseus, Cambridge, MA, (2002).

[4] A.-L. Barabasi, R. Albert, Emergence of scaling in random networks, *Science,* **286** (1999) pp.509-512.

[5] A.-L. Barabasi, R. Albert, H. Jeong, Mean-field theory for scale-free random networks, *Phys. A,* **272** (1999) pp.173-187.

[6] A.-L. Barabasi, R. Albert, H. Jeong, Mean-field theory for scale-free random networks: The topology of the World Wide Web, *Physics A.* **281** (2000) pp.69-77.

[7] A.-L. Barabasi, R. Albert, H. Jeong, G. Bianconi, Power-law distribution of the World Wide Web, *Phys. A,* **287** (2000) pp.2115a.

[8] K. Bharat, B. Cheng, M. Henzinger, M. Ruhl, Who Links to Whom: Mining Linkage between Web Sites, *Proceedings of the IEEE International Conference on Data Mining,* (2001) pp.51-58.

[9] B. Bollobas, Mathematical results on scale-free random graphs, *Handbook of graphs and networks,* Wiley-VCH, Weinheim, (2003), pp.1-34.

[10] B. Bollobas, O. Riordan, The diameter of a scale-free graph, *preprint,* Department of Methematical Sciences, University of Memphis, (2002).

[11] B. Bollobas, O. Riordan, J. Spencer, G. Tusnady, The degree sequence of a scale-free random graph process, *Random Structures Algorithms,* **18** (2001) pp.279-290.

[12] S. Boyd, G. Labonte, Finding the exact integrality gap for small traveling salesman problems, *IPCO 2002* (2002) pp.83-92.

[13] A. Broder, R. Kumar, F. Maghoul, P. Raghava, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structures in the web, *Computer Networks* **33** (2000) pp.309-320.

[14] R. Carr, R. Ravi, A new bound for the 2-edge connected subgraph problem, *IPCO'98* (1998).

[15] T. Christof, A. Lobel, M. Stoer, PORTA: A POlyhedron Representation Transformation Algorithm, *http://www.zib.de/Optimization/Software/Porta/index.html* (1997)

[16] N. Christofides, Worst case analysis of a new heuristic for the traveling salesman problem, *Report 388, Graduate School of Industrial Administration,* Carnegie Mellon University, Pittsburgh (1976).

[17] J. Edmonds, Edge-disjoint brankings, *Combinatorial Algorithms* (1973) pp.91-96

[18] P. Erdos, A. Renyi, On random grpahs I., *Pulicationes Mathematicae Debrecen* **5** (1959) pp.290-297.

[19] P. Erdos, A. Renyi, On evolution of random graphs, *Magyar Tud. Akad. Mat. Kutao Int. Kozl.* **5** (1960) pp.17-61.

[20] Google homepage: http://www.google.com, Accessed in April 2005.

[21] L. Lovasz, Combinatorial Problems and Exercises, North-Holland, Amsterdam, (1979).

[22] T. Luczak, P. Pralat, Protean Graphs, *Internet Mathematics,* (accepted) (2004)

[23] T. Luczak, P. Pralat, Protean Graphs - Giant Component and its Diameter, *Internet Mathematics,* (submitted) (2005)

[24] H. M. Mahmoud, R. T. Smythe, J. Szymanski, On the structure of random plane-oriented recursive trees and their branches, *Random Structures and Algorithms,* **4** (1993) pp.151-176.

[25] B. McKay, nauty User's Guide (Version 1.5), *Technical Report TR-CS-90-02*, Department of Computer Science, Australia National University, (1991).

[26] M. E. J. Newman, The Structure and Function of Complex Networks *Siam Review,* **45** (2003) pp.167-256.

[27] B. Pittel, Note on the heights of random recursive trees and random m-ary search trees, *Random Structures and Algorithms,* **5** (1994) pp.337-347.

[28] P. Pralat, Grafy proteuszowe [Protean Graphs (In Polish)], *Ph.D. dissertation,* (2004)

[29] D. J. de S. Price, Networks of scientific papers, *Science,* **149** (1965) pp.510-515.

[30] D. J. de S. Price, A general theory of bibliometric and other cumulative advantages processes, *Journal of the American Society for Information Science,* **27** (1976) pp.292-306.

[31] D. B. West, *Introduction to Graph Theory*, Prentice-Hall, Upper Saddle River, NJ, (1996)