INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700  800/521-0600
AN ASYMPTOTIC STUDY OF SEVERAL MODELS OF SPARSE RANDOM GRAPHS

DISSertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Robert Stephen Weishaar, M.S., M.A.S.

* * * * *

The Ohio State University
1997

Dissertation Committee:
Professor Boris Pittel, Adviser
Professor Thomas Dowling
Professor Gerald Edgar

Approved by

Boris Pittel
Adviser

Department of Mathematics
ABSTRACT

Some asymptotic properties of several models of sparse random graphs are studied. First, the performance of the greedy coloring algorithm “first fit” on random graphs $G_{n,c/n}$ is analyzed. It is shown that approximately $\log_2 \log n$ colors are used, the exact number being concentrated with high probability (whp) on at most two consecutive integers. We obtain a similar result for random trees. Next, we study a process for generating random 2-regular multigraphs with loops, called the 2-hook process, which we show produces a graph with cycle lengths distributed according to the Ewens sampling formula (with parameter 1/2) from population genetics. We obtain distributions on the lengths of the first cycles formed by the process. Finally, we study connectivity and matching properties of the bipartite $k$th nearest neighbor graphs, $B_k$. We show that the expected matching number of $B_1$ is approximately 0.8. For $B_2$, we show that whp the graph is either connected or consists of a giant component and several cycles with bounded total size. We derive a formula for the probability that the graph is connected and show that whp the graph does not have a perfect matching. For $B_k$, $k \geq 3$, we show that the graph is connected and has a perfect matching whp.
ACKNOWLEDGMENTS

I wish to thank my adviser, Professor Boris Pittel, for all of his time and encouragement. He made this dissertation possible, but more importantly, he made the time spent completing it enjoyable. I greatly appreciate his concern for me not only as a developing mathematician, but as a person as well.

I am thankful for the support of Professor Anant Godbole, Professor Vladimir Tonchev, Professor Alphonse Baartmans and Professor Don Kreher from Michigan Technological University. It was with them that I had my first experiences with mathematical research.

I appreciate Professor Thomas Dowling, Professor Gerald Edgar and Professor Robert Bartoszynski taking time out of their schedules to serve on my dissertation committee.

I thank my office mates, especially Tom, George and Kendra, for their friendship and support throughout the process.

I am grateful for my church family, in particular the members of my small groups and the Senior High youth group members and other sponsors for the wonderful ways in which they have encouraged me.
I thank my parents for their continual support, and for taking an interest in my studies. The same can be said of the rest of my family, including my wife's entire family.

Finally, I thank my wife, Robin, for her encouragement and patience. She did not simply allow me to focus on my studies; she was usually the one doing the focusing. There would be no dissertation without her.

This document was typeset using \textsf{AMSTeX}.
VITA

April 12, 1970 .......... Born - Milwaukee, Wisconsin

1992 .................... B.S. Mathematics, Michigan Technological University

1992-1993 ............ Graduate Teaching Assistant,
                      Michigan Technological University

1993 .................... M.S. Mathematics, Michigan Technological University

1997 .................... M.A.S., The Ohio State University

1993-present ........ Graduate Teaching and Research Associate,
                      The Ohio State University

PUBLICATIONS


FIELDS OF STUDY

Major Field: Mathematics
TABLE OF CONTENTS

Abstract ............................................................................................................................ ii
Acknowledgments ........................................................................................................... iii
Vita .................................................................................................................................... v

Chapters:
1. Introduction ............................................................................................................... 1
2. On line coloring of sparse random graphs ............................................................... 8
   2.1 Introduction ...................................................................................................... 8
   2.2 Coloring $G_{n,c/n}$ ............................................................................................. 9
   2.3 Coloring random trees ................................................................................... 18
3. The 2-hook process ................................................................................................. 23
   3.1 Introduction .................................................................................................... 23
   3.2 Cyclic structure of $H_n$ .................................................................................. 25
   3.3 Formation of first cycles ............................................................................... 28
4. The bipartite nearest neighbor graphs $B_k$ ............................................................ 46
   4.1 Introduction .................................................................................................... 46
   4.2 Matching number of $B_1$ .............................................................................. 50
4.3 Properties of $B_2$ ......................................................... 65
4.4 Properties of $B_k$, $k \geq 3$ ............................................. 84
4.5 Open problems ............................................................... 89

List of References ............................................................... 91
CHAPTER 1

INTRODUCTION

Formally, a graph $G$ is a pair $(V, E)$, where $V$ is a set of vertices and $E$ is a set of unordered pairs of elements from $V$, called edges. We say that two vertices are adjacent iff the pair is an edge of $G$. If edges may be repeated the structure is called a multigraph, and if vertices may be adjacent to themselves we call the structure a graph (or multigraph) with loops. Usually the term graph by itself implies that there are no multiple edges or loops, though for the sake of brevity we may use the term more loosely when the actual structure is clear from the context.

The theory of random graphs was founded by Erdős and Rényi [11,12]. The books by Palmer [28] and Bollobás [4] contain many of the important results from this fascinating topic. The theory rests on the following definition from probability theory. A probability space is a triple $(\Omega, \Sigma, P)$, where $\Omega$ is a set, $\Sigma$ is a $\sigma$-field of subsets of $\Omega$ and $P$ is a non-negative measure on $\Sigma$ with $P(\Omega) = 1$. For our purposes, $\Omega = \Omega_n$ will consist of a finite set of graphs on vertex set $V_n = \{1, 2, \ldots, n\}$, and $\Sigma$ will be the power set of $\Omega$. 
Perhaps the most natural model of a random graph is when \( \Omega_n \) is the set of all graphs on \( n \) vertices, where each graph is equally likely. Notice that there are \( N = \binom{n}{2} \) potential edges in a graph with \( n \) vertices, and each graph corresponds to a particular subset of these potential edges. Therefore, each graph occurs with probability \( 2^{-N} \).

The two most common models of random graphs are denoted \( G(n, M) \) and \( G(n, p) \), where \( M \) is usually and \( p \) is often a function of \( n \). The first model consists of all graphs with vertex set \( V_n \) that have \( M \) edges, where each graph is equally likely. \( G(n, M) \) has \( \binom{N}{M} \) elements so each occurs with probability \( \binom{N}{M}^{-1} \). The second model, \( G(n, p) \), with \( 0 < p < 1 \), consists of all graphs with vertex set \( V_n \) in which each of the \( N \) potential edges are chosen independently and with probability \( p \). If \( G_0 \) is a graph on \( V_n \) with \( m \) edges, then

\[
P(\{G_0\}) = P(G = G_0) = p^m q^{N-m}
\]

where \( q = 1 - p \). In general, \( G \) will denote a random graph from \( G(n, M) \), \( G(n, p) \) or any other model we define (the particular model will be clear from the context).

Notice that \( p = 1/2 \) in \( G(n, p) \) corresponds to the model in which every graph on \( n \) vertices is equally likely.

A third common model of random graph is the random regular graph. This model, denoted \( G(n, r - \text{reg}) \) assigns equal probability to each \( r \)-regular graph on \( n \) vertices (where \( r n \) is even).

In the study of random graphs, it is usually the case that we have a sequence of probability spaces, one for each \( n \), and we are interested in properties of the space.
as \( n \to \infty \). We use whp (with high probability) to indicate that the probability of an event is \( 1 + o(1) \) as \( n \to \infty \).

One very active area of research in graph theory has been graph coloring. In fact, one of the most famous mathematics problems of all time is whether a map can be colored with four colors in such a way that no two countries which share a border receive the same color. In graph theoretical terms, the four-color theorem asserts that every loopless planar graph has a vertex 4-coloring. Conjectured by Guthrie in 1852, this theorem was finally proved by Appel and Haken [1,2,3] in 1976. Robertson, Sanders, Seymour and Thomas [35] provide a simpler proof, using the same basic approach as Appel and Haken.

With graph coloring being such an important topic, it is no surprise that much work has been done with the coloring of random graphs. The chromatic number of a graph is the fewest number of colors that must be used to color the vertices in such a way that no two adjacent vertices receive the same color. Results on the chromatic number for the random graph \( G(n,p) \) are given in Grimmet and McDiarmid [14], Bollobás and Erdős [6], Shamir and Spencer [38], Bollobás [5] and Luczak [23]. In general, to color a graph with the fewest number of colors possible requires a view of the entire graph. In some applications, however, you might be given the vertices one at a time, and must color them immediately ("on-line"). Certainly you are likely to use more colors in this situation. The interest is in finding an algorithm that doesn't do too badly. The greedy algorithm "first fit" is a popular choice and has been analyzed for several different classes of graphs. Available is a list of colors,
and each vertex is colored with the first admissible color on the list. Together [14] and [5] show that for \( G(n, p) \), \( p \) fixed, first fit uses only about twice as many colors as an optimal coloring would.

In Chapter 2 we apply the algorithm first fit to \( G(n, \frac{c}{n}) \) and to random trees. We show that the number of colors required is approximately \( \log_2 \log n \) (compared to a chromatic number which has been shown to be bounded in probability (see [4])) and that the number of colors used is concentrated on at most 2 or 3 integer values respectively whp, similar to Luczak's [23] result that the chromatic number is concentrated on 2 values for \( G(n, p) \), \( p < n^{-5/6-\varepsilon} \), whp.

Two important methods used in our proof for the concentration of the number of colors used are the first and second moment methods. (They are also used extensively in Chapter 4.) If \( X \) is a nonnegative integer-valued random variable, then \( P(x \geq 1) \leq E[X] \). Notice that if \( E[X] = o(1) \), then \( X = 0 \) whp. And, from Chebyshev's inequality,

\[
P(X = 0) \leq \frac{\text{var}[X]}{E[X]^2}.
\]

In particular, (1.1.1) implies that if \( E[X] \to \infty \) and \( \text{var}[X] = o(E[X]^2) \), then \( X > 0 \) whp.

Depending on the model, sampling from our probability spaces may be a difficult task. \( G(n, p) \) is particularly easy to work with in many situations, mainly due to the independence between the edges. In \( G(n, m) \), the edges are not independent — the presence of one edge makes the presence of another less likely. In this situation,
it is often convenient to think of the graph as evolving in time. Beginning with the empty graph on \( n \) vertices, \( m \) edges are sequentially inserted. Sampling from \( \mathcal{G}(n, r - \text{reg}) \) is much more difficult. Bollobás (see [4]) gives a model which makes the study of random regular graphs accessible if \( r \) does not grow too quickly with \( n \). Because generating a graph with all vertices of degree \( r \) uniformly is difficult, often simpler algorithms are used to generate regular graphs, with a loss in uniformity. One such algorithm is called the \( d \)-process and is analyzed by Ruciński and Wormald [36,37]. Starting with an empty graph on \( n \) vertices, edges are randomly added one by one so that the degree of each vertex remains bounded by \( d \). In [36] they show that if \( nd \) is even, then whp the \( d \)-process produces a regular graph. In [37] they study the 2-process in greater detail, showing some differences between probabilities for a graph produced by the 2-process and a 2-regular graph chosen uniformly at random. Another algorithm for generating \( d \)-regular graphs is the star \( d \)-process, analyzed by Robalewska [33] and Robalewska and Wormald [34]. In this process the vertices are caused to have degree \( d \) one after the other.

A third natural approach to generating regular graphs is to initially endow each vertex with a capacity for exactly \( d \) edges. Starting with an empty graph on \( n \) vertices, \( d \) hooks are connected to each vertex and at each step two hooks are selected at random and an edge connects the corresponding vertices. Notice that loops and multiple edges are possible in this model. This model, called the \( d \)-hook process, seems particularly natural in an application in which the restriction on the vertex degrees is a result of some property of the vertices themselves, or if the less
encumbered vertices have a greater tendency to "react" with the others. In chapter 3 we study the 2-hook process. We show that the final cyclic structure of the graph has a distribution given by the Ewens sampling formula (with parameter 1/2) from population genetics and mention several results arising from this relationship. Then we use a differential equations approach from Pittel, Spencer and Wormald [31] to keep track of the number of isolated vertices remaining in the graph in order to derive a distribution for the sizes of the first cycles formed in the process.

Consider a random graph process where edges are inserted one at a time at random. The $d$-process studied by Rucinski and Wormald is a restriction of this in which only edges which do not increase the degree of either end vertex above $d$ are included. If instead we only require that not both end vertices have degree increased above $k$ we obtain a model called the $k^{th}$ nearest neighbor graph, denoted $O_k$. The name comes from an equivalent model in which each edge of the complete graph $K_n$ has a length coming from some continuous distribution. Any edge that is one of the $k$ shortest incident with either end vertex is included in the graph. This model is studied by Cooper and Frieze in [8] where they are primarily concerned with the connectivity of these graphs.

In Chapter 4, we study the bipartite version of the nearest neighbor graphs, denoted $B_k$, derived from a complete bipartite graph with random lengths $C_{i,j}$, $1 \leq i, j \leq n$. As in the case of the nonbipartite graphs, we show that $B_2$ is either connected or consists of several cycles of bounded total size and a giant component whp, and that $B_k$, $k \geq 3$, is connected whp.
One problem associated with the complete bipartite graph and the random lengths (or costs) of the edges which underlies our graph is the random costs assignment problem which seeks to find the cost of a minimal perfect matching. One method of attacking the problem is to find suboptimal matchings by looking at graphs with reduced size. $B_k$ is a natural choice for the reduced graph. We prove that $B_3$ has a perfect matching whp. We also prove that $B_2$ does not have a perfect matching whp, but conjecture that the portion of vertices matched tends to 1 as $n \to \infty$, as suggested by computer experiments. The conditioning device, found useful in the study of the stable marriage problem (see Pittel [30]), is used extensively in our work.

Continued study of these graphs should lead to a better understanding of the assignment problem and may lead to better bounds on the cost of a minimum matching, a problem studied by Walkup [42], Karp [18], Dyer, Frieze and McDiarmid [10] and Karp, Rinnoy Kan and Vohra [19], among others.
2.1 Introduction.

The chromatic number of a graph $G$, denoted $\chi(G)$, is the fewest number of colors that can be used to color the vertices of $G$ in such a way that no two adjacent vertices receive the same color. No polynomial time algorithm is known that finds the chromatic number of a graph. Here we are interested in the performance of the algorithm first fit for coloring certain types of graphs. To color a graph using this greedy algorithm one simply chooses the vertices of the graph at random, coloring each with the smallest color still available, i.e. with the smallest color that has not been used on any neighbor already colored. We let $\chi_g(G)$ denote the number of colors used by an application of the greedy algorithm to $G$. Optimally, a sparse random graph can be colored with a bounded number of colors whp. This may, however, be difficult to do — or impossible, if we cannot see the whole graph and are only given the vertices in an on-line fashion. See [27] for a broad discussion of on-line algorithms, and [15] and [17] for an investigation of on-line coloring of other classes of graphs. We show that first fit uses about $\log_2 \log n$ colors whp for $G_{n,c/n}$ and random trees. We also show that the number of colors used by first fit
is concentrated on at most two colors whp for $G_{n,c/n}$ complementing similar results of Shamir and Spencer [38] and Luczak [23] for the chromatic number itself. For random trees, the number of colors used is concentrated on at most three values whp. We remark that first fit is easy to apply, and that even for very large values of $n$, $\log_2 \log n$ is very reasonable.

2.2 Coloring $G_{n,c/n}$.

Though first fit rarely gives an optimal coloring, i.e. uses only $\chi(G)$ colors, we may hope that it doesn’t do too badly. Consider the case of $G_{n,p}$. When $p$ is fixed, this greedy algorithm does quite well. In fact, Grimmet and McDiarmid [14] have shown that

$$\frac{(1-\epsilon)n}{\log_b n} \leq \chi(G) \leq \frac{(1+\epsilon)n}{\log_b n} \text{ whp,}$$

and that

$$\chi(G) \geq \frac{\left(\frac{1}{2} - \epsilon\right)n}{\log_b n} \text{ whp,}$$

for any fixed $\epsilon$, where $b = \frac{1}{1-p}$.

Later Shamir and Spencer [38] showed that the chromatic number of random graphs (with any $p$) is sharply concentrated on at most $n^{1/2} \omega(n)$ consecutive integers, where $\omega(n)$ goes to infinity arbitrarily slowly. Bollobás [5] then showed

$$\frac{\left(\frac{1}{2} - \epsilon\right)n}{\log_b n} \leq \chi(G) \leq \frac{\left(\frac{1}{2} + \epsilon\right)n}{\log_b n} \text{ whp,}$$

and thus first fit uses about twice as many colors as an optimal coloring would.
However, when $p$ is small, say $p = c/n$, first fit doesn't perform so well. In fact, it is well known that (see [4])

$$\chi(G_{n,c/n}) \leq \ell \quad \text{whp.}$$

(2.2.1)

where $\ell$ is a constant depending only on $c$, but for all fixed $k$,

$$\chi_g(G_{n,c/n}) > k \quad \text{whp.}$$

(2.2.2)

For an easy demonstration of (2.2.2), consider the sequence of trees $T_k$ defined recursively as follows. $T_1 :=$ a single vertex. $T_k := T_{k-1}$ with each vertex acquiring a new pendant vertex. By calculating moments and applying Chebyshev's inequality, it is straightforward to show that for each fixed $k$, $G_{n,c/n}$ contains $\Omega(n)$ components isomorphic to $T_k$ whp. Notice that if the vertices of $T_k$ are colored "from the outside in", $k$ colors will be required. With so many copies of $T_k$ present in the sparse random graph, the order in which the vertices are encountered when applying the greedy algorithm will result in the use of color $k$ for one of them whp.

For smaller values of $p$, Shamir and Spencer [38] have a sharper concentration result. They showed that when $p < n^{-5/6-\epsilon}, \epsilon > 0$, the chromatic number of $G(n, p)$ is whp concentrated on at most five consecutive integers. Luczak [23] improved this to show $\chi(G)$ takes on one of two values whp. Our main result shows that the number of colors used by first fit to color $G(n, c/n)$ also takes on one of two values whp and indicates how fast $\chi_g(G)$ grows with $n$. 
Theorem 2.2.1. Let $G_{n,c/n}$ be a random graph. Then,

$$\log_2 \log n + j(c) - f_1(c) < \chi_g(G_{n,c/n}) < \log_2 \log n + j(c) - f_2(c) \quad \text{whp},$$

where $f_1(c)$, $f_2(c)$ and $j(c)$ depend only on $c$, and where

$$j(c) = \min\{j : c_j \leq \frac{1}{2}\},$$

where $c_1 = c$ and for $j \geq 2$, $c_j$ is given recursively by

$$c_j = c_{j-1} - \log(c_{j-1} + 1).$$

Moreover, $\sup(f_1(c) - f_2(c)) \leq 1.5$, so that the number of colors used is concentrated on at most two values whp.

To see what the underlying issues are, let's examine the greedy algorithm at work, applied to $G_{n,c/n}$. From Pittel [29] (cf. Bollobás [4], Chvátal [7]), the first color class has about $\frac{\log(c+1)}{c} n$ vertices whp. If we remove the vertices of the first color class, the remaining graph has a random number of vertices $N$ such that

$$\left| \frac{cN}{n(c - \log(1 + c))} - 1 \right| < \epsilon$$

with high probability. The probability of an edge between a pair of vertices in the resulting graph remains the same, namely $p = c/n$. (After all, no such edge has been exposed yet.) In other words, conditioned on $N = n'$, the resulting random graph is distributed like $G_{n',c/n}$. 

11
In terms of $n'$ we have

$$p = \frac{c'}{n'}, \quad c' = \frac{cn'}{n},$$

and for the likely range of $n'$, $c' \sim c - \log(c + 1)$.

If $c \geq \frac{1}{2}$, $\log(c + 1) \geq \log\frac{3}{2}$ and so for such $c$ there is a constant $j(c)$ depending only on $c$ such that removing the first $j(c)$ color classes results in a graph on $N^*$ (random) vertices, where, conditioned on $N^* = n^*$, this graph is distributed like $G_{n^*, c^*/n^*}$, and for the likely range of $n^*$, $c^* < \frac{1}{2}$.

To estimate the number of colors used we do the following. Remove the first $j(c)$ color classes, then make use of two properties of random graphs $G_{n, c/n}$ with $c < 1$: Each component is either a tree or unicyclic whp (this result implies (2.2.1)), and the number of vertices on unicyclic components is bounded in probability. For proofs, see Bollobás [4]. (While these properties require only $c^* < 1$, removing $j(c)$ color classes to get $c^* < \frac{1}{2}$ allows us to get a stronger concentration result.) Shamir and Upfal [39] introduce a graph coloring algorithm which begins by applying this greedy algorithm and is followed by an optimal coloring of the graph that remains when removing color classes results in a graph whose components are trees or unicyclic whp. We are interested in the number of colors used when we are not permitted a view of the remaining graph, and must continue to color on-line. We have already seen that $\chi_g(G)$ is unbounded in probability. Since the number of vertices on unicyclic components is bounded in probability, the components requiring $\chi_g(G)$ colors are trees whp.
Proposition 2.2.2. When applying the greedy algorithm to a graph $G$, a tree component will receive at least $k$ colors iff it contains a subtree $S_k$ with the following properties:

(i) $S_k$ is isomorphic to $T_k$,

(ii) The pendant vertices of $S_k$ receive color 1. Vertices which become pendant subsequent to the removal of the pendant vertices receive color 2, etc., and

(iii) When looked at apart from the component in which it lies, the vertices of $S_k$ are colored in an order which forces the use of color $k$.

Proof. If $k$ colors are required, there must be a vertex receiving color $k$. It must be connected to, and colored after, a vertex of color $k - 1$. Each of these vertices must be connected to, and colored after, vertices of color $k - 2$. Continuing until we get to vertices receiving color 1 gives us a tree that has properties (i), (ii) and (iii). Conversely, if one of these subtrees is present, $k$ colors are used. □

Apply first fit to $G_{n,c/n}, c < \frac{1}{2}$. Let $B_k$ (for 'bad') denote the number of trees satisfying properties (i), (ii) and (iii). (Note that (ii) and (iii) are not equivalent, nor does one imply the other.) $\chi_g(G_{n,c/n}) \geq k$ iff $B_k \geq 1$. Finding $B_k$ is not an easy task; (ii) causes the problem. We instead consider two other quantities which bound $B_k$. Let $B'_k$ (for 'bad and lonely') denote the number of isolated trees satisfying (i), (ii) and (iii). (For isolated trees, (ii) and (iii) are equivalent.) Let $A_k$ (for 'alright') denote the number of trees satisfying only (i) and (iii). Then $B'_k \leq B_k \leq A_k$. 

13
To estimate these quantities we need to calculate some moments. To do this we need to find the number of labeled graphs on $2^{k-1}$ vertices satisfying (i) and (iii).

**Proposition 2.2.3.** For $k \geq 2$, the number of distinct labeled graphs on $2^{k-1}$ vertices isomorphic to $T_k$ is $\frac{1}{2} \left(2^{k-1}\right)!$.

**Proof.** When $k = 2$, $T_k$ is an edge on two vertices. There is only one such graph. Assume the formula is true for $k = m - 1$. We show it is then true for $k = m$.

For $m > 2$, $T_m$ consists of a set of pendant vertices and, after removing these, a graph isomorphic to $T_{m-1}$. To count the number of ways this may happen, we choose $2^{m-2}$ vertices on which the subgraph isomorphic to $T_{m-1}$ will sit, count the number of these, then multiply by the number of ways the pendant vertices may be paired with the vertices in this subgraph. We get

\[
\binom{2^{m-1}}{2^{m-2}} \cdot \frac{1}{2} \left(2^{m-2}\right)! \cdot \left(2^{m-2}\right)! = \frac{1}{2} \left(2^{m-1}\right)!.
\]

**Proposition 2.2.4.** For $k \geq 2$, the probability that $T_k$ will require $k$ colors when its vertices are colored at random is $\frac{1}{2^{2^{m-1}-2}}$.

**Proof.** When $k = 2$, $T_k$ is an edge and requires two colors. Assume the formula is true for $k = m - 1$. We show it is then true for $k = m$. Each pendant vertex of $T_m$ must be colored before its neighbor, and the set of neighbors, inducing a subgraph isomorphic to $T_{m-1}$, must be colored in a proper order, which has probability given by the induction hypothesis. We get:

\[
\frac{1}{2^{2^{m-2}}} \cdot \frac{1}{2^{2^{m-2}-2}} = \frac{1}{2^{2^{m-1}-2}}.
\]
Let's begin by calculating $E(B'_k)$, using $t$ for the number of vertices on each of these trees, i.e. $t = 2^{k-1}$.

$$E(B'_k) = \binom{n}{t} \left( \frac{c}{n} \right)^{t-1} \left( 1 - \frac{c}{n} \right)^{t+1+t(n-t)} \cdot \frac{\frac{1}{2} t!}{2^{t-2}}$$

where $\rho = 2/c$. For the values of $k$ we will be considering, the approximation we use for the binomial coefficient, i.e.,

$$\binom{n}{t} \sim \frac{n^t}{t!},$$

is valid. In particular, this approximation is valid for $t = O(\log n)$.

Let's explain where the factors in (1.3) come from. We take a subset of vertices of size $t$, then multiply by the probability that an isolated tree sits on these vertices. Notice there must be edges between $t - 1$ pairs of the vertices with no edges between the remaining $\binom{t}{2} - (t - 1)$ pairs of vertices, and no edges between the $t(n - t)$ pairs of vertices with one vertex in the tree, the other outside the tree. Then we multiply by the number of trees having the desired properties — the number of trees on $t$ vertices isomorphic to $T_k$ times the probability that one requires $k$ colors.

We also need $E[(B'_k)^2]$. Since we are working with isolated trees, this is not difficult. We compute $E(B'_k(B'_k - 1))$ by finding the probability that on an ordered pair of distinct sets of $t$ vertices there lie trees with the desired properties, and
multiply by the number of ways the sets may be chosen.

\[
E(B_k'(B_k' - 1)) = \binom{n}{t} \left(\frac{n - t}{t}\right) \left(\frac{c}{n}\right)^{2(t-1)} \\
\cdot \left(1 - \frac{c}{n}\right)^{2\binom{t}{2} - 2t + 2 + 2t(n - 2t) + t^2} \left(\frac{\frac{1}{2}t!}{2^{t-2}}\right)^2
\]

Thus if \( E(B_k') \to \infty \) we have \( E((B_k')^2) \sim [E(B_k')]^2 \), and so \( \text{Var}(B_k') \) is \( o([E(B_k')]^2) \).

We conclude that if \( E(B_k') \to \infty \) then \( B_k' \to \infty \) whp, using Chebyshev's inequality.

We can bound \( E(A_k) \) from above by ignoring the structure of the components on which each of the trees contributing to \( A_k \) lies, i.e. ignoring that each is on a tree component. For each set of \( t \) vertices we take the probability there is a tree on these vertices and multiply by the number of trees with the desired properties. We obtain

\[
E(A_k) \leq \binom{n}{t} \left(\frac{c}{n}\right)^{t-1} \left(1 - \frac{c}{n}\right)^{(t)} (t - 1) - t + 1 \frac{1}{2^{t-2}}
\]

\[
\sim \frac{nct^{-1}}{2^{t-1}} = \frac{n}{\rho^{t-1}}.
\]

Now suppose \( 2^{k-1} = t > \frac{(1+\epsilon)\log n}{\log \rho} \). Then

\[
\log E(A_k) \leq \log n - \frac{(1 + \epsilon)\log n}{\log \rho} \log \rho + \log \rho
\]

\[
= -\epsilon \log n + \log \rho \to -\infty
\]

and thus \( B_k = 0 \) whp. In other words, \( k \) colors would not be used whp. We obtain an upper bound for the number of colors used:

\[
\chi_{\rho}(G) \leq \log_2 \log n - \log_2 \log \rho + 1 \quad \text{whp.} \quad (2.2.5)
\]
If $2^{k-1} = t < \frac{(1-\varepsilon)\log n}{\log \rho + c}$ then

$$
\log E(B_k') \geq \log n - c \frac{(1-\varepsilon)\log n}{\log \rho + c} - \frac{(1-\varepsilon)\log n}{\log \rho + c} \log \rho + \log \rho
$$

$$
= \varepsilon \log n + \log \rho \to \infty
$$

and thus $B_k > 0$ whp. That is, $k$ colors are used whp. This gives us a lower bound on the greedy chromatic number:

$$
\chi_g(G) \geq \log_2 \log n - \log_2(\log \rho + c) \quad \text{whp.}
$$

The proof of Theorem 2.2.1 is complete, with

$$
f_1(c) = \log_2(\log \frac{2}{c^*} + c^*), \quad \text{and}
$$

$$
f_2(c) = \log_2(\log \frac{2}{c^*} - 1),
$$

where $c^* = c_j(c)$.

Our analysis provides some insight into the performance of first fit on $G_{n,m}$. Consider the following: (1) It can be shown that $G_{n,c/n}$ has $\left\lceil \frac{cn}{2} \right\rceil$ edges with probability $\Omega(n^{-1/2})$. (2) By taking $\varepsilon = 1$ in (2.2.4) we obtain $E(A_k) = O(n^{-1})$, and thus $P(k$ colors are required) $= O(n^{-1})$. Since, conditioned on the number of edges $= m$, $G_{n,p}$ is equally likely to be any graph with $m$ edges, (1) and (2) yield: If $G$ is randomly chosen (uniformly) from the set of all graphs with $\left\lceil \frac{cn}{2} \right\rceil$ edges, then, analogous to (2.2.5),

$$
\chi_g(G) \leq \log_2 \log n - \log_2(2/c) + 2 \quad \text{whp.}
$$
2.3 Random Trees.

The chromatic number of a tree is 2, as with any bipartite graph. A simple random algorithm will successfully color any tree with two colors:

1. Select a vertex at random, assign it color 1.
2. Select at random an uncolored vertex from the set of vertices adjacent to colored vertices. Color it with the smallest color available.
3. Repeat (2) until all vertices have been colored.

Simply changing the selection rule to include all vertices, rather than only those with a neighbor that has been colored, drastically changes the performance of the algorithm. The modified algorithm is first fit, the one used earlier to color $G_{n,c/n}$ (see section 2.1), and as was the case there, the number of colors used to color a random tree is unbounded in probability. Our next result indicates how fast the number of colors needed grows.

**Theorem 2.3.1.** Let $T$ be a random tree on $n$ vertices. Then

$$\log_2 \log n - c_1 \leq \chi_g(T) \leq \log_2 \log n + c_2, \quad \text{whp},$$

where $c_1 = .76$ and $c_2 = 1.53$.

Let $T_k$, $B_k$, and $A_k$ be defined as before. There are no isolated subtrees in a random tree, so we modify the definition of $B'_k$ to let us get a lower bound for $\chi_g(T)$. Let $B'_k$ be the number of subtrees isomorphic to $T_k$ with only one edge connecting it to any other vertex of the random graph, where that vertex is colored after any
vertex of the tree isomorphic to $T_k$, and the vertices of the tree are colored in an order which requires the use of $k$ colors.

Let's begin by computing $E(A_k)$, the expected number of subtrees isomorphic to $T_k$ so that, when looked at apart from the rest of the tree, the vertices are colored in an order which requires the use of $k$ colors. By Cayley's formula (see Moon [26]), for each set of $t$ vertices, there are $tn^{n-t-1}$ forests of $t$ trees, each tree containing one of those $t$ vertices. By Proposition 2.2.3 there are $\frac{1}{2}t!$ trees on the $t$ vertices isomorphic to $T_k$. Multiplying these quantities and dividing by the total number of trees on $n$ vertices gives us the probability that a random tree $T$ has a subtree isomorphic to $T_k$ on a given set of $t$ vertices. We then multiply by the probability — given by Proposition 2.2.4 — that the vertices are colored in an appropriate order.

$$E(A_k) = \frac{n}{t} \cdot \frac{tn^{n-t-1}}{n^{n-2}} \cdot \frac{1}{2^{t-2}}.$$
the vertices of the tree on $t$ vertices are colored in an appropriate order.

$$E(B'_k) = \binom{n}{t} \frac{1}{n^{n-2}} \frac{\frac{1}{t!}(n-t)^{n-t-2}t(n-t)}{t+1} \frac{1}{2t-2}$$

$$\sim \frac{n^t(n-t)^{n-t-1}}{n^{n-2}2^{t-1}} \sim \frac{n^t}{2t}.$$

We also need $E[(B'_k)^2]$. Since each of the trees contributing to $B'_k$ has only one attachment to the rest of $T$, no two of the trees share a vertex. We need only consider whether each in the pair is attached to the same vertex. For each pair of sets of $t$ vertices with no vertices in common we compute the probability that each satisfies the required properties to obtain

$$E(B'_k(B'_k - 1)) = \binom{n}{t} \binom{n-t}{t} \frac{1}{n^{n-2}} \frac{(n-2t)^{n-2t-2}}{t}$$

$$\cdot \left(\frac{(n-2t)(n-2t-1)t^2}{2^{2t-4}(t+1)^2} + \frac{(n-2t)t^2}{2^{2t-4}(2t+1)}\right)$$

$$\sim \frac{n^t(n-t)^t(n-2t)^{n-t-2}}{2^{2t-4}} \frac{n^t}{2^{2t-4}} \sim \frac{n^te^{-2t}}{2^{2t-2}}.$$

The two terms inside the parentheses in the second line represent two cases. If the two trees are attached to different vertices, there are $(n-2t)(n-2t-1)t^2$ ways to choose attachment vertices, and we multiply by the probability the vertices are colored in an appropriate order. If they are attached to the same vertex, there are only $(n-2t)t^2$ ways to choose the vertices, and the common attachment must be colored after all $2t$ vertices, which has probability $1/(2t+1)$. The first term is clearly dominant.

If $E(B'_k) \to \infty$ we have $E[(B'_k)^2] \sim [E(B'_k)]^2$, which implies, using Chebyshev's inequality, that $B'_k \to \infty$ whp.
Suppose $2^{k-1} = t > \frac{(1+\varepsilon)\log n}{\log 2}$. Then, since $\log t - (t - 1) \log 2$ is decreasing for $t > (\log 2)^{-1}$, we have

$$\log(\mathbb{E}(A_k)) \leq \log n + \log \frac{(1 + \varepsilon)\log n}{\log 2} - (1 + \varepsilon)\log n + \log 2$$

$$\sim -\varepsilon \log n \to -\infty.$$

Thus $B_k = 0$ whp, so $k$ colors are not needed whp. We obtain the following upper bound:

$$\chi_g(T) \leq \log_2 \log n - \log_2 \log 2 + 1 \quad \text{whp.}$$

Now suppose $2^{k-1} = t < \frac{(1-\varepsilon)\log n}{\log 2+1}$. Then

$$\log(\mathbb{E}(B_k')) \geq \log n - \frac{(1 - \varepsilon)\log n}{\log 2 + 1} - \frac{(1 - \varepsilon)\log n}{\log 2 + 1} \log 2$$

$$\sim \varepsilon \log n \to \infty.$$

Thus $B_k' \to \infty > 0$ whp, whence $B_k > 0$ whp, so $k$ colors are needed whp. This gives us a lower bound:

$$\chi_g(T) \geq \log_2 \log n - \log_2(\log 2 + 1) \quad \text{whp.}$$

The proof of Theorem 2.1 is complete. We remark that in the case of random trees, the greedy coloring algorithm uses one of at most 3 possible numbers of colors whp.

It is interesting to compare the performance of another common greedy algorithm often used to color graphs. This algorithm begins with a set $\{1, 2, \ldots, K(n)\}$ of
colors and, when given a vertex to color, randomly chooses one from the subset of admissible colors. The question is, given $K(n)$, will we successfully color the graph (whp)? Consider a random tree $T$. It is well known that the maximum degree of a random tree is $\sim \frac{\log n}{\log \log n}$ whp (see Moon [26]). Let $x$ denote the vertex with maximum degree. The neighbors of $x$ colored before $x$ receive colors that are chosen independent of one another (since we're working with a tree). If $f(n) = o(\deg(x))$, $x$ will be colored after $f(n)$ of its neighbors whp. Also, if $f(n) \gg K(n) \log K(n)$, all $K(n)$ colors will be used on a set of $g(n)$ independently colored vertices whp (the coupon collecting problem). With $K(n) = \lceil \log n/((\log \log n)^2 + \epsilon) \rceil$, $\epsilon > 0$ and $f(n) = \lceil \log n/((\log \log n)^{1+\epsilon}/2) \rceil$ we see this greedy algorithm whp fails to color the tree using $K(n)$ colors, and thus compares unfavorably with first fit. The same is true for sparse random graphs, since there we find an equally high degree vertex on some isolated trees whp.
CHAPTER 3
THE 2-HOOK PROCESS

3.1 Introduction

Because of the difficulty of generating regular graphs uniformly at random (see [24]), there is interest in simpler algorithms which produce regular graphs with a distribution that is not perfectly uniform. Ruciński and Wormald study such an algorithm. The $d$-process forms a graph of maximum degree $d$ by randomly inserting edges between vertices as long as that edge doesn’t make one of the vertices have degree greater than $d$. In [36] they show that whp the graph formed has $\lfloor nd/2 \rfloor$ edges, and so if $nd$ is even a $d$-regular graph is formed whp. In [37] they study the $2$-process in greater detail.

Another algorithm of this type is the star $d$-process, defined and analyzed by Robalewska and Wormald [34], in which a $d$-regular graph is also formed with high probability. Robalewska [33] studies the star $2$-process in greater detail. In the case of the star process, the vertices are sequentially saturated.

While looking at the work done by Ruciński and Wormald, we were led to consider a similar process, which we call the $d$-hook process, for generating a $d$-regular
multigraph with loops on \( n \) vertices, where \( dn \) even. Start with an empty graph on \( n \) vertices. Label \( dn \) balls, \( d \) for each vertex of the graph. Randomly select two balls and join the corresponding vertices with an edge, which may be a loop or multiple edge. Continue the process until all the balls are drawn. Alternatively, we can think of the vertices of the empty graph as each having \( d \) places of attachment, called hooks, and at each step two hooks are randomly selected. This method of generation we call the \( d \)-hook process, and the resulting \( d \) regular multigraph with loops we denote by \( H_n^d \). Here we are concerned with the \( 2 \)-hook process and let \( H_n = H_n^2 \).

The \( 2 \)-hook process results in a graph with the \( n \) vertices partitioned into a number of cycles. Most of the interesting properties of the final graph concern the sizes of the cycles formed, rather than which vertices are on which cycles or the order of the vertices on a particular cycle. Thus, we are mainly interested in the sizes of the parts in the partition of the \( n \) vertices. Let \( \Pi_n \) denote the set of partitions of \( n \). It is conventional to represent \( \pi \in \Pi_n \) in the form \( \pi = 1^{\alpha_1}2^{\alpha_2} \cdots n^{\alpha_n} \), where \( \alpha_i \) is the number of parts of size \( i \). For example, \( \{1, 1, 1, 2, 3, 3, 5\} \) (a partition of 16) would be written as \( 1^32^13^25^1 \). The Ewens sampling formula, originally arising in genetics [13], defines a distribution on \( \Pi_n \). The Ewens sampling formula with parameter \( \theta > 0 \) assigns probability

\[
\frac{n!}{\theta(\theta + 1) \cdots (\theta + n - 1)} \prod_{i=1}^{n} \frac{\theta^{\alpha_i}}{i^{\alpha_i} \alpha_i!}
\]

(3.1.1)

to \( \pi \).
Our first theorem is that the 2-hook process partitions the \( n \) vertices into cycles with sizes distributed according to the Ewens sampling formula with parameter \( \theta = 1/2 \). Section 2 contains the proof of this assertion. We also mention some of the properties of \( H_n \) that follow from this relationship with the Ewens sampling formula.

There is also interest in how the graph evolves. When does the first cycle form? How big are the first cycles? In section 3 we focus our attention on the evolution of the process. Our main theorem concerns the distribution of the length of the \( m^{th} \) cycle formed by the 2-hook process, where \( m \) is fixed.

### 3.2 Cyclic structure of \( H_n \)

We begin by discussing some properties of the 2-hook process that will be important in this section and the next. After the insertion of the \( j^{th} \) edge (time \( j \)), the graph consists of several cycles (possibly) and \( n - j \) components which are paths, some of which are simply isolated vertices. There are exactly \( 2(n - j) \) hooks remaining. The probability that \( e \) is the next edge to appear is given by

\[
\frac{w(e)}{\binom{2(n-j)}{2}},
\]

where \( w(e) \) is a probability weight equal to 0, 1, 2 or 4 depending on the number of hooks free on the two vertices involved. For example, if one vertex is isolated and the other is of degree one, then there are two pairs of hooks that result in \( e \) appearing, thus \( w(e) = 2 \).
The probability that a cycle is formed with the insertion of the next edge is given by

\[
\frac{n - j}{\binom{2(n-j)}{2}} = \frac{1}{2(n-j)-1},
\]

regardless of the history of the process, since the two hooks chosen must be the free pair from one of the \(n-j\) non-cyclic components.

Property (3.2.2) allows us to answer a number of questions very quickly. The expected number of cycles is easily seen to be \(\sim \frac{1}{2} \log n\).

We can also compute the probability that the resulting graph is a cycle of length \(n\). It is simply \(P(A)\), where \(A\) is the event that no cycle is formed until the insertion of the last edge. Using Stirling's approximation we get the asymptotic probability of the event \(A\):

\[
P(A) = \prod_{j=1}^{n-1} \frac{2(n-j)}{2(n-j)+1} = \frac{(2n-2)!!}{(2n-1)!!} \sim \frac{\sqrt{\pi}}{2\sqrt{n}}.
\]

The next theorem will allow us to say much more about the structure of \(H_n\).

**Theorem 3.2.1.** The probability that \(H_n\) contains \(\alpha_i\) cycles of length \(i\) for \(i = 1, 2, \ldots, n\) is given by the Ewens sampling formula (3.1.1) with \(\theta = 1/2\).

**Proof.** Let \(P_x\) denote the probability in question. Initially, we will also consider the order in which the edges appeared and find the product of the probability weights for the edges in the graph. Notice that a loop has a probability weight of 1. Now consider a cycle of length \(k+1 > 1\). The last edge inserted has a probability weight 1. Discarding this edge, we have a path of length \(k\). We can show that each
appearance order for its edges is equally likely, and equal, in terms of the probability weights, to $2^{k+1}$. This is true if the path has length 1. For a longer path, consider the last edge inserted. If it is on the end, it has a weight of 2, and the first $k - 1$ edges have a weight product of $2^k$, by induction, so the claim is true. If the last edge is in the middle, it has a weight factor of 1, and it connects paths of length $r$ and $k - r - 1$. The product of the probability weights is $2^{r+1} \cdot 2^{k-r-1+1} = 2^{k+1}$, by induction. Thus, a cycle of length $k > 1$ has a probability weight of $2^k$. Therefore, our graph has a probability weight of $2^{n-\alpha_1}$. Notice that this value does not depend on the order in which the edges appeared. The number of processes leading to this graph is $n! / 2^{\alpha_2}$, since there are $n$ edges which may be arranged in any order, but each cycle of length 2 consists of a pair of identical edges. Now we must calculate the number of graphs with the same cyclic structure. A well known formula originally due to Cauchy gives the number of permutations of $n$ with $\alpha_i$ cycles of length $i$ as

$$n! \prod_{i=1}^{n} \frac{1}{i^{\alpha_i} \alpha_i !}.$$  

We simply divide this expression by $\prod_{i=3}^{n} 2^{\alpha_i}$, since for $i \geq 3$ there are two cyclic permutations of $i$ elements which lead to the same cycle in the graph. Combining factors, we obtain

$$P_n = \prod_{j=0}^{n-1} \binom{2(n-j)}{2}^{-1} \cdot 2^{n-\alpha_1} \cdot \frac{n!}{2^{\alpha_2}} \prod_{i=1}^{n} \frac{1}{i^{\alpha_i} \alpha_i !} \prod_{i=3}^{n} 2^{\alpha_i}$$

$$= \frac{2^n}{(2n)!} 2^n n! \prod_{i=1}^{n} \frac{(1/2)^{\alpha_i}}{i^{\alpha_i} \alpha_i !}$$

$$= \frac{n!}{\frac{1}{2} \left(\frac{1}{2} + 1\right) \cdots \left(\frac{1}{2} + n - 1\right)} \prod_{i=1}^{n} \frac{(1/2)^{\alpha_i}}{i^{\alpha_i} \alpha_i !}.$$

$\Box$
Let $K_n : [0, 1] \times \Pi_n \rightarrow \mathbb{R}$ be defined by $K_n(u, \pi) = \text{the number of parts in } \pi \text{ of length at most } n^u$ for $\pi \in \Pi$. Hansen [16] proves the following.

**Theorem 3.2.2.** If $\pi$ has a distribution given by the Ewens sampling formula (3.1.1), then as $n \to \infty$ the random element

$$Y_n(u, \pi) = \frac{K_n(u, \pi) - \theta u \log n}{\sqrt{\theta \log n}}$$

converges weakly to Brownian motion on $[0, 1]$.

In particular, this theorem implies that the number of cycles in $H_n$ converges to a normal distribution with mean and variance $\frac{1}{2} \log n$.

Kingman [20, 21] has shown that when written in decreasing order and normalized by $n$, the parts of a partition with distribution (3.1.1) converge to the Poisson-Dirichlet distribution with parameter $\theta$. Also, the joint distribution of the number of parts of sizes 1, 2, ..., $k$, for any fixed $k$, converges to that of independent Poisson random variables with means $\theta, \theta/2, \ldots, \theta/k$, respectively.

### 3.3 Formation of first cycles.

We can prove the following concerning the time at which the first cycles are formed. Let $T_m$ be the time at which the $m^{th}$ cycle is formed. Since the process is complete at time $n$, we write $\tau_m = T_m/n$ and obtain:

**Proposition 3.3.1.** The limiting density function for the proportion of time passed when the $m^{th}$ cycle is formed is

$$f_m(x) = f_{\tau_m}(x) = \frac{(-1)^{m-1} \log^{m-1}(1 - x)}{2^m (m - 1)! \sqrt{\log(1 - x)}}.$$
Proof. We have

\[
P(\tau_1 > x) = P(T_1 > xn)
\]

\[
= \prod_{j=1}^{[xn]} \left(1 - \frac{1}{2n - 2j + 1}\right)
\]

\[
\sim \exp \left(-\frac{1}{2} \sum_{j=1}^{[xn]} \frac{1}{n - j}\right)
\]

\[
= \exp \left(-\frac{1}{2} \sum_{j=[(1-x)n]}^{[n]} \frac{1}{j}\right)
\]

\[
\sim \exp(-\frac{1}{2}(\log n - \log(1 - x)n))
\]

\[
= \exp(\frac{1}{2} \log(1 - x))
\]

\[
= \sqrt{1 - x},
\]

from which we obtain \( f_1(x) = \frac{1}{2\sqrt{1-x}} \).

Now assume the proposition is true for the \((m-1)\)th cycle. If \(\tau_m-1 = y\), then \(\tau_m - y\) is distributed like \(\tau_1^{1-y}\), the time of formation of the first cycle in a process on \([(1-y)n]\) vertices. Thus,

\[
P(\tau_m > x|\tau_m-1 = y) = P(\tau_m - y > x - y|\tau_m-1 = y)
\]

\[
= P(\tau_1^{1-y} > x - y)
\]

\[
= \sqrt{1 - \frac{x - y}{1 - y}}
\]

\[
= \frac{\sqrt{1 - x}}{\sqrt{1 - y}},
\]

from which we obtain,

\[
f_{\tau_m|\tau_m-1}(x|y) = \frac{1}{2\sqrt{1-y}\sqrt{1-x}}.
\]
We can then compute

\[
    f_m(x) = \int_0^x f_{r_m | r_{m-1}}(x|y) f_{m-1}(y) dy
\]

\[
    = \int_0^x \log^{m-2}(1 - y)^{-1} \frac{1}{2^{m-1}(m-2)! \sqrt{1 - y} 2 \sqrt{1 - y} \sqrt{1 - x}} dy
\]

\[
    = \frac{(-1)^{m-2}}{2^m(m-2)!} \int_0^x \frac{\log^{m-2}(1 - y)}{(1 - y) \sqrt{1 - x}} dy
\]

\[
    = \frac{(-1)^{m-1}}{2^m(m-2)!} \int_0^x \log^{1-x} \frac{u^{m-2}}{\sqrt{1 - x}} du
\]

\[
    = \frac{(-1)^{m-1} \log^{m-1}(1 - x)}{2^m(m-1)! \sqrt{1 - x}}. \quad \square
\]

Let's investigate the size of various cycles. The resulting graph can be described by a list of the hooks in the order they were chosen. We imagine the process complete, and expose the cycles in the reverse order in which they were formed. Each hook is equally likely to have been chosen last. The last hook could have been connected to any other. If this second hook lies on the same vertex as the first, then the last cycle formed has length 1. The probability of this is \( \frac{1}{2n-1} \). If it's on a different vertex, take the other hook on that vertex and find the hook to which it was connected. Continue until a cycle is formed. Notice that the time until the cycle is formed is distributed as the time of formation of the first cycle in the 2-hook process, but here the time also corresponds to the length of the cycle! Thus the distribution of the length (as a fraction of \( n \)) of the last cycle is given by

\[
    g(x) = \frac{1}{2^{m-1} \sqrt{1 - x}}.
\]

After the last cycle formed has been exposed, we proceed in the same manner to expose the second to last cycle formed, and so on. The length of the second to last cycle formed is distributed as the time between formation of the
first and second cycles in the original forward process. In general, the lengths of the last cycles formed are distributed like differences between consecutive random variables distributed according to the sequence of distributions given in Proposition 3.3.1.

Let's consider the first cycle formed. For the first cycle to have length \( \ell \), the following must happen. As we're exposing the cycles in reverse order, at the point when there are \( 2\ell + 2 \) hooks left the next one selected must be connected to the hook which completes a cycle. This has probability \( \frac{1}{2\ell + 1} \). From now on, the hook which completes the cycle may not be chosen, until there is only one hook left. Thus the probability that the first cycle has length \( \ell \) is given (exactly, for all \( n \)) by

\[
\frac{1}{2\ell + 1} \frac{2\ell - 2}{2\ell - 1} \cdots \frac{4}{5} = \frac{(2\ell - 2)!!}{(2\ell + 1)!!}
\]

for \( \ell < n \). The probability that the first cycle has length \( n \) is the probability that the process results in a single cycle, which is given by (1.1). Let \( L \) be the length of the first cycle formed. Then we can compute

\[
E(L) = \sum_{\ell=1}^{n-1} \ell \cdot \frac{(2\ell - 2)!!}{(2\ell + 1)!!} + n \cdot \frac{(2n - 2)!!}{(2n - 1)!!} \sim \frac{\sqrt{\pi n}}{2} + \frac{\sqrt{\pi n}}{2} = \sqrt{\pi n}.
\]

We have been able to find the probability distribution for the length of the first cycle formed. Trying to find the same for subsequent cycles leads to some complicated summation formulas based on conditioning on the lengths of earlier cycles. In the hope of finding a relatively simple expression for the probability distribution of the length of the \( m^{th} \) cycle, we abandon this method of exposing the
cycles in reverse order and concentrate on the forward process, making use of the
following:

(1) If at time \( j \) we know how many vertices remain (that is, are not yet on a cycle) and how many paths (of length at least one) have been formed, we can calculate the probability that a randomly chosen path has \( \ell \) vertices.

(2) The first few cycles contain few vertices whp.

(3) We can estimate the number of paths in the graph with sufficient accuracy.

We formalize each of these by proving a number of lemmas.

**Lemma 3.3.2.** Let \( G \) be a random graph on \( n \) vertices defined by the 2-hook process at time \( j \). Conditioned on the absence of cycles and \( \{ \) the number of paths of length at least one \( = b \} \), \( G \) is uniformly distributed on the set of all graphs with \( n \) vertices, \( j \) edges, no cycles, and \( b \) paths of length at least one. Furthermore, when \( b > 1 \), the distribution of the number of vertices, \( X \), on a randomly chosen path is given, for \( \ell \geq 2 \), by

\[
P(X = \ell) = \frac{(j-\ell)}{(b-1)}.
\]

*Proof.* First we show uniformity. As demonstrated in the proof of Theorem 3.2.1, the probability weight for a path on \( k \) edges is \( 2^{k+1} \). Thus the probability weight of any graph with the given conditions is

\[
\prod_{i=1}^{b} 2^{k_{i}+1} = 2^{\sum(k_{i}+1)} = 2^{j+b}.
\]

Since this expression depends only on \( j \) and \( b \), \( G \) is conditionally uniform.
To find the probability distribution of the length of a randomly chosen path, let's calculate the number of graphs with the given conditions. We must choose \( j + b \) vertices to belong to the paths, partition the vertices into \( b \) unordered sets corresponding to the \( b \) paths with the condition that each set has at least two vertices, and finally multiply by the number of nonisomorphic labelings. We obtain the following:

\[
\binom{n}{j + b} \frac{(j - 1)!}{b!} \frac{(j + b)!}{2^b}.
\]

Now, the probability that we select a particular path with \( \ell \) vertices is given by \( 1/b \) times the number of graphs satisfying the given conditions which contain this path, divided by (3.3.2). Then, we multiply by the number of distinct paths with \( \ell \) vertices. The numerator of this expression is

\[
\frac{1}{b} \binom{n - \ell}{j + b - \ell} \binom{j - \ell}{b - 2} \frac{1}{(b - 1)!} \frac{(j + b - \ell)!}{2^{b-1}} \binom{n}{\ell} \frac{\ell!}{2}.
\]

After dividing by (3.3.2) we obtain the desired result. □

A probabilistic upper bound on the number of vertices belonging to the first few cycles is given by the following lemma.

**Lemma 3.3.3.** Let \( A_m \) be the event that the first \( m \) cycles use a total of more than \( \sqrt{n} \) vertices. Then, for fixed \( m \),

\[
P(A_m) = O\left(\frac{\log^m n}{n^{1/4}}\right).
\]

**Proof.** We imagine the cycles formed sequentially as implied by the reverse process.
which led to (3.3.1), and use induction. Then,

\[
P(A_1) = \frac{1}{2[\sqrt{n}] - 1} \cdot \frac{2[\sqrt{n}] - 4}{2[\sqrt{n}] - 3} \cdot \frac{4}{5} + \frac{1}{2[\sqrt{n}] + 1} \cdot \frac{2[\sqrt{n}] - 2}{2[\sqrt{n}] - 1} \cdot \frac{4}{5} + \ldots
\]

\[
\leq \sum_{j=1}^{n} \frac{1}{2j - 1} \cdot \frac{1}{2^m - 1} \cdot \frac{(2[\sqrt{n}] - 4)!}{(2[\sqrt{n}] - 3)!} + O\left(\frac{\log n}{n^{1/4}}\right)
\]

\[
\sim \frac{1}{4} \log n \frac{\sqrt{\pi}}{2n^{1/4}} = O\left(\frac{\log n}{n^{1/4}}\right).
\]

Assume the lemma is true for the first \(m-1\) cycles. If we let \(k_i\) be the time when the \(i\)th cycle is formed, and \(K = \{k_1, \ldots, k_{m-1} : 0 < k_1 < \cdots < k_{m-1} < \sqrt{n}\}\), then

\[
P(A_m) = P(A_{m-1}^c, A_m) + O\left(P(A_{m-1}^c)\right)
\]

\[
\leq \sum_{j=1}^{n} \frac{1}{2j - 1} \cdot \sum_{k=1}^{m-1} \frac{1}{2^m - 1} k_{m-1} \cdot \frac{(2[\sqrt{n}] - 4)!}{(2[\sqrt{n}] - 3)!} + O\left(\frac{\log n}{n^{1/4}}\right)
\]

\[
\leq \sum_{j=1}^{n} \frac{1}{2j - 1} \cdot \prod_{i=1}^{m-1} \frac{1}{2j - 1} k_{m-1} \cdot \frac{(2[\sqrt{n}] - 4)!}{(2[\sqrt{n}] - 3)!} + O\left(\frac{\log n}{n^{1/4}}\right)
\]

\[
= O\left(\frac{\log n}{n^{1/4}}\right) \quad \square
\]

We will need a concentration result for \(b_j\), the number of paths of length at least one at time \(j\), for a process conditioned on the absence of cycles. After the insertion of edge \(j\) there are \(2(n - j)\) hooks remaining. If there are \(b_j\) paths, they account for \(2b_j\) hooks. The remaining \(2(n - j - b_j)\) hooks are on isolated vertices, thus there are \(i_j = n - j - b_j\) isolated vertices at time \(j\). So we can concentrate on obtaining a concentration result for \(i_j\).

We follow an approach of Pittel, Spencer and Wormald [30]. The idea is to write a differential equation system for the expectations of parameters of the graph process,
and then use martingale techniques to show that the actual values of the parameters are close to the solution of the system whp. See Rhee and Talagrand [32] for an application of martingale inequalities to other combinatorial problems, and see Talagrand [40,41] for the ‘majorizing measures’ approach of bounding parameters of stochastic processes. For a graph $G$, we let $w(G) = (i(G), \mu(G))$, where $i(G)$ is the number of isolated vertices of $G$ and $\mu(G)$ is the number of edges of $G$. If $G(j)$ is the graph obtained by the $2$-hook process at time $j$, then we write $w_j = w(j) = w(G(j))$, with $i_j$ and $\mu_j$ defined analogously. Let $W$ be the set of graphs with no cycles. Then we write

$$P(w_{j+1} = (i - a, j + 1) | w_j = (i, j), G(j + 1) \in W) = p(w_{j+1} | w_j).$$

Then,

$$p(w_{j+1} | w_j) = \begin{cases} \frac{(2(n-j-i)) - (n-j-i)}{(2(n-j)) - (n-j)}, & a = 0 \\ \frac{2(n-j-i)(2i)}{(2(n-j)) - (n-j)}, & a = 1 \\ \frac{(2i) - i}{(2(n-j)) - (n-j)}, & a = 2 \\ \frac{(n-j-i)(n-j-i-1)}{(n-j)(n-j-1)}, & a = 0 \\ \frac{2(n-j-i)i}{(n-j)(n-j-1)}, & a = 1 \\ \frac{i(i-1)}{(n-j)(n-j-1)}, & a = 2 \\ \frac{(n-j-i)^2}{n-j} + O \left( \frac{1}{n-j} \right), & a = 0 \\ \frac{2(n-j-i)i}{(n-j)^2} + O \left( \frac{1}{n-j} \right), & a = 1 \\ \frac{i}{n-j} + O \left( \frac{1}{n-j} \right), & a = 2 \end{cases}$$

(3.3.3)
Similarly define

\[ q(w_{j+1}|w_j) = \begin{cases} \left(\frac{n-j-i}{n-j}\right)^2 & a = 0 \\ \frac{2(n-j-i)i}{(n-j)^2} & a = 1 \\ \left(\frac{i}{n-j}\right)^2 & a = 2 \end{cases} \]

Then,

\[ E_q(i_{j+1} - i_j| i_j = i) = -\frac{2i(n-j-i)}{(n-j)^2} - 2\left(\frac{i}{n-j}\right)^2 = \frac{-2i}{n-j}, \]

and

\[ E_q(\mu(j + 1) - \mu(j)|\mu(j)) = 1. \]

Letting \( x = \frac{i}{n} \), and \( y = \frac{j}{n} \), we get the following system of approximate differential equations for the means \( E[w(j)] \):

\[ x' = \frac{-2x}{1-y}, \]

\[ y' = 1. \]

With the initial conditions \( x(0) = 1 \) and \( y(0) = 0 \), we obtain \( x = (1-y)^2 \).

Let

\[ J(w_j) = x - (1-y)^2 = \frac{ij}{n} - \left(1 - \frac{j}{n}\right)^2. \]

**Lemma 3.3.4.** For all \( u \) and \( \alpha < \min(1/2, u) \),

\[ P\left\{ \max_{j \leq n-n^u} |J(w_j)| > n^{-\alpha} \right\} = O\left(e^{-n^{2-\alpha}}\right). \]
Proof. Let

\[ Q(w_j) = \exp\{n^\alpha J(w_j)\}. \]

Then

\[ Q(w_{j+1}) = Q(w_j) \exp\{n^\alpha [J(w_{j+1}) - J(w_j)]\}. \]

Now,

\[ J(w_{j+1}) - J(w_j) = \left( \frac{i_{j+1}}{n} - \left(1 - \frac{j + 1}{n}\right)^2 \right) - \left( \frac{i_j}{n} - \left(1 - \frac{j}{n}\right)^2 \right) \]

\[ \quad = \frac{i_{j+1} - i_j}{n} + O(n^{-1}) = O(n^{-1}), \]

so that,

\[ J(w_{j+1}) = J(w_j) + (w_{j+1} - w_j) \cdot \text{grad } J(w_j) + O(n^{-2}). \]

Therefore,

\[ Q(w_{j+1}) = Q(w_j) \left[ 1 + n^\alpha (w_{j+1} - w_j) \cdot \text{grad } J(w_j) + O(n^{2(\alpha - 1)}) \right]. \]

Then,

\[ \mathbb{E}[Q(w_{j+1})|w_j] = Q(w_j) \left[ 1 + n^\alpha \mathbb{E}[(w_{j+1} - w_j) \cdot \text{grad } J(w_j)] + O(n^{2(\alpha - 1)}) \right]. \]

Now,

\[ \mathbb{E}[J(w_{j+1}) - J(w_j)|w_j] \cdot \text{grad } J(w) = \sum_{w_{j+1}} (w_{j+1} - w_j) \cdot \text{grad } J(w_j) \]

\[ \quad \cdot \mathbb{E}_q[w_{j+1}|w_j] - q(w_{j+1}|w_j), \]

with the \( q \) term allowed because \( J \) is constant along the trajectory of the differential equation system derived from \( q \) (3.3.4), so that

\[ \mathbb{E}_q[w_{j+1} - w_j|w_j] \perp \text{grad } J(w_j). \]
Now, from (3.3.3) and (3.3.4),

$$|p(w_{j+1}|w_j) - q(w_{j+1}|w_j)| = O((n - j)^{-1}),$$

and from (3.3.5),

$$\text{grad } J(w_j) = O(n^{-1})$$

so that

$$E[J(w_{j+1}) - J(w_j)|w_j] = O((n(n - j))^{-1}).$$

Therefore,

$$E[Q(w_{j+1})|i] = Q(w_j) \left( 1 + O \left( \frac{n^\alpha}{n(n - j)} \right) + O \left( \frac{n^{2\alpha}}{n^2} \right) \right)$$

$$= Q(w_j)(1 + O(n^{-\omega})).$$

where \(\omega > 1\), because \(\alpha < \min(1/2, \nu)\). Then the sequence

$$\{R(j)\} := \{(1 + n^{-\omega} \log n)^{-j} Q(w_j)\}$$

is a supermartingale.

Introduce a stopping time

$$\mathcal{T} = \begin{cases} 
\min\{j < n - n^u : J(w_j) > n^{-d}\}, & \text{if such a } t \text{ exists,} \\
n - n^u, & \text{otherwise.}
\end{cases}$$

Let \(\mathcal{T}' = \min\{\mathcal{T}, [n - n^u]\}\). Then, applying the Optional Sampling Theorem (see Durrett [4]) to \(\{R(j)\}\) and \(\mathcal{T}'\), we get, in terms of \(\{Q(w_j)\}\),

$$E[Q(w_{\mathcal{T}'})] \leq (1 + n^{-\omega} \log n)^n \cdot E[Q(w_0)]$$

$$= (1 + n^{-\omega} \log n)^n = O(1).$$

38
Now if $T \leq [n - n^u]$, then $J(w_T) = n^{-d}$, and so $Q(w_T) = e^{n^{a-d}}$, thus

\[ E[Q(w_T)] \geq e^{n^{a-d}} \cdot P(T \leq [n - n^u]), \]

so that

\[ P\left\{ \max_{j \leq n-n^u} J(w_j) > n^{-d}\right\} = P(T \leq [n - n^u]) = O(e^{-n^{a-d}}). \]

Analogously,

\[ P\left\{ \max_{j \leq n-n^u} J(w_j) < -n^{-d}\right\} = O(e^{-n^{a-d}}). \]

In light of the fact that we do not have a strong concentration result for the number of isolated vertices toward the end of the process, we prove the following.

**Lemma 3.3.5.** Let $D_m$ be the event that the $m^{th}$ cycle is formed after time $n-n^{2/3}$. Then

\[ P(D_m) = O\left(\frac{\log^{m-1} n}{n^{1/6}}\right). \]

**Proof.** $D_1$ is simply the probability that no cycles are formed before time $n-n^{2/3}$. Modifying (3.2.3), we obtain

\[ P(D_1) = \prod_{j=1}^{n-[n^{2/3}]} \frac{2(n-j)}{2(n-j) + 1} = O(n^{-1/6}). \]

Now let $D'_m$ be the event that exactly $m$ cycles are formed before time $n-n^{2/3}$. Let $k_i$ be the time when cycle $i$ is formed, and let

\[ K = \{k_1, \ldots, k_m : 0 < k_1 < \cdots < k_m < n-n^{2/3}\}. \]
Then,

\[ P(D'_m) = \sum_{K} \frac{1}{2^{m_1} \cdot k \cdot \ldots \cdot k_m} \cdot \prod_{j=1}^{n-\lfloor n^{2/3} \rfloor} \frac{2(n-j)}{2(n-j)+1} \]

\[ \leq \prod_{i=1}^{m} \sum_{k_i=1}^{n} \frac{1}{k_i} \cdot \prod_{j=1}^{n-\lfloor n^{2/3} \rfloor} \frac{2(n-j)}{2(n-j)+1} \]

\[ = O \left( \frac{\log^m n}{n^{1/6}} \right). \]

Then we have

\[ D_m = D'_0 + D'_1 + \ldots + D'_{m-1} = O \left( \frac{\log^m n}{n^{1/6}} \right). \]  

**Lemma 3.3.6.** Let \( T_k^{n'} \) denote the time of formation of the \( k \)th cycle and \( L_k^{n'} \) denote the number of vertices belonging to the first \( k \) cycles formed for a 2-hook process defined on \( n' \) vertices, with the superscript suppressed when \( n' = n \). Then we have:

\[ P(L_{k+1} = L_k + \ell, T_{k+1} = t | L_k = \ell', T_k = t') \]

\[ = P(L_{1}^{n-\ell'} = \ell, T_{1}^{n-\ell'} = t - \ell' | T_{1}^{n-\ell'} > t' - \ell'), \] \hspace{1cm} (3.3.6)

\[ P(L_1 = 1 | T_1 = t + 1, b_t = b) = \left( 1 - \frac{b}{n-t} \right), \] \hspace{1cm} (3.3.7)

\[ P(L_1 = \ell | T_1 = t + 1, b_t = b) = \frac{(t-\ell)}{(t-1)} \frac{b}{n-t} \] \hspace{1cm} (3.3.8)

if \( \ell > 1 \), and

\[ P(T_{1}^{m-n} = t - m | T_{1}^{m-n} > t' - m) = P(T_1 = t | T_1 > t'). \] \hspace{1cm} (3.3.9)

**Proof.** (3.3.6) follows from a general principle. Suppose the vertices of the graph are partitioned into two sets, \( D \) and \( E \), with no edges between them at time \( t \).
The graphs on $D$ and $E$ have evolved like 2-hook processes on $|D|$ and $|E|$ vertices respectively.

For (3.3.7), notice that a loop is formed iff one of the $i_t = n - t - b_t$ pairs of hooks on an isolated vertex are chosen of the $n - t$ pairs of hooks which complete a cycle.

To prove (3.3.8), notice that if a cycle of length $\ell > 1$ is formed at time $t$ then the hooks at the end of a path on $\ell$ vertices were selected. Since each pair of hooks have the same chances of being selected, Lemma 1 gives the correct probability, if we condition on a path of length at least 1 being selected. The extra factor in the expression removes this condition. We note that this expression is the same for all $n' > t$.

Finally, recall that the probability that a cycle is formed at time $j$ for a 2-hook process defined on $n'$ vertices is $(2(n' - j) + 1)^{-1}$. (3.3.9) follows from the fact that this expression depends only on the difference $n' - j$, and not on how $n'$ or $j$ individually varies. □

Lemma 3.3.7. If $t < n - n^{2/3}$, then

$$P(L_1 = \ell|T_1 = t + 1) = \frac{t^{\ell-1}(n-t)}{n^\ell} + O(n^{-1/9}).$$  (3.3.10)

Proof. From lemma 3.3.4,

$$P\left(\max_{t \leq n - n^{2/3}} |i_t - n(1 - t/n)^2| > n^{5/9}\right) = O(e^{-n^\delta}),$$  (3.3.11)

where $\delta < 1/18$. Thus, from (3.3.7) and since $b_t + i_t = n - t$,

$$P(L_1 = 1|T_1 = t + 1, b_t) = \frac{i_t}{n - t}$$
so that
\[
P(L_1 = 1|T_1 = t + 1) = \frac{n(1 - t/n)^2 + O(n^{5/9})}{n - t} + O(e^{-n^4}) = 1 - t/n + O(n^{-1/9}).
\]

From (3.3.8), for \( \ell > 1 \),
\[
P(L_1 = \ell|T_1 = t + 1, b_t) \leq \frac{b_t}{n - t}.
\]

Since \( b_t \leq t \),
\[
P(L_1 = \ell|T_1 = t + 1) \leq t/(n - t) = O(n^{-1/6})
\]
if \( t \leq n^{5/6} \). For larger values of \( t \), we simplify (3.3.8) and get
\[
P(L_1 = \ell|T_1 = t + 1, b_t = b) = \frac{(b - 1) \left( (t - b)^{\ell - 2} + O((t - b)^{\ell - 3}) \right)}{t^{\ell - 1} + O(t^{\ell - 2})} \cdot \frac{b}{n - t}.
\]

Now from (3.3.11),
\[
b = (n - t) - n(1 - t/n)^2 + O(n^{5/9})
\]
and
\[
t - b = t - (n - t) - n(1 - t/n)^2 + O(n^{5/9})]
\[
= t^2/n + O(n^{5/9}),
\]
with probability \( 1 - O(e^{-n^4}) \). After substituting for \( b \) and \( t - b \) and simplifying, we obtain
\[
P(L_1 = \ell|T_1 = t + 1) = \frac{t^\ell(n - t)}{n^\ell} + O \left( \frac{n^{5/9}}{t} \left( \frac{t}{n} \right)^{\ell - 2} \right).
\]

With \( t > n^{5/6} \), the error term is \( O(n^{-5/18}) \), which finishes the proof. \( \square \)
Theorem 3.3.8. Let $G$ be a random graph defined by the 2-hook process. Let $p_{k,\ell}$ be the limiting probability that the $k^{th}$ cycle has length $\ell$. Then for all fixed $k \geq 1$ and $\ell \geq 1$,

$$p_{k,\ell} = \sum_{m=0}^{\ell-1} \binom{\ell-1}{m} \frac{(-1)^m}{(2m+3)^k}.$$

Proof. We begin by rewriting

$$p_{k+1,\ell} = P(L_{k+1} - L_k = \ell) = \sum_{t' = k}^{n-\ell} \sum_{t \geq t'} g(\ell, t, \ell', t'),$$

where

$$g(\ell, t, \ell', t') = P(L_{k+1} - L_k = \ell, T_{k+1} = t+1, L_k = \ell', T_k = t').$$

Then we write

$$g(\ell, t, \ell', t') = P(L_{k+1} - L_k = \ell, T_{k+1} = t+1 | L_k = \ell', T_k = t') \cdot P(L_k = \ell', T_k = t'),$$

and making use of (3.3.6) and some further conditioning,

$$g(\ell, t, \ell', t') = P(L_1^{n-\ell'} = \ell, T_1^{n-\ell'} = t+1 - \ell' | T_1^{n-\ell'} > t' - \ell') \cdot P(L_k = \ell', T_k = t')$$

$$= P(L_1^{n-\ell'} = \ell | T_1^{n-\ell'} = t+1 - \ell') \cdot P(T_1^{n-\ell'} = t+1 - \ell' | T_1^{n-\ell'} > t' - \ell') \cdot P(L_k = \ell', T_k = t').$$

Because of lemmas 3.3.3 and 3.3.5, we can write

$$p_{k+1,\ell} = \sum_{\ell' = k}^{\lfloor n^{1/2} \rfloor} \sum_{t = 1}^{\lfloor n^{3/2} \rfloor} \sum_{t' = 1}^{t} g(\ell, t, \ell', t') + o(n^{-1/7}).$$
With these limits on the variables, using lemma 3.3.7,

\[ P(L_1^{n-t'} = \ell | T_1^{n-t'} = t + 1 - \ell') = \frac{(t - \ell')^{t-1}(n - t + \ell')}{n^t} + O(n^{-1/9}) = \frac{t^{t-2}(n - t)}{n^t} + O(n^{-1/2}) + O(n^{-1/9}). \]

Now (3.3.8) provides

\[ g(\ell, t, \ell', t') = \left( \frac{t^{t-2}(n - t)}{n^t} + O(n^{-1/9}) \right) \cdot P(T_1 = t | T_1 > t') \cdot P(L_k = \ell', T_k = t'). \]

Only the last factor involves \( \ell' \) now. Using the concentration result of lemma 3.3.3 again we can write

\[ p_{k+1,t} = \sum_{t=1}^{[n-n^{2/3}]} \sum_{t'=1}^{t} \frac{t^{t-1}(n - t)}{n^t} \cdot P(T_1 = t | T_1 > t') \cdot P(T_k = t') + O(n^{-1/9}). \]

Now letting \( x = t/n \) and \( x' = t'/n \), letting \( n \) go to infinity, and using Proposition 3.3.1, we get

\[ p_{k+1,t} = \int_0^1 \int_0^x x^{x-1}(1-x) \frac{1}{2\sqrt{1-x}\sqrt{1-x'}} \frac{(-1)^{k-1} \log^{k-1}(1-x')}{2^k \sqrt{1-x'(k-1)!}} \, dx' \, dx \]

\[ = \frac{(-1)^{k-1}}{2^{k+1}(k-1)!} \int_0^1 \int_0^x x^{x-1}\sqrt{1-x} \log^{k-1}(1-x') \, dx' \, dx \]

\[ = \frac{(-1)^{k}}{2^{k+1}k!} \int_0^1 x^{x-1}\sqrt{1-x} \log^k(1-x) \, dx \]

\[ = \frac{(-1)^{k}}{2^{k+1}k!} \int_0^1 (1-x)^{x-1}\sqrt{x} \log^k x \, dx. \quad (3.3.12) \]
Then, by the Binomial Theorem,

\[
p_{k+1,t} = \frac{(-1)^k}{2^{k+1}k!} \int_0^1 \sum_{m=0}^{\ell-1} \binom{\ell}{m} (-1)^m x^m \log^k x \, dx
\]

\[
= \frac{(-1)^k}{2^{k+1}k!} \sum_{m=0}^{\ell-1} \binom{\ell}{m} (-1)^m \int_0^1 x^m \log^{k+1} x \, dx
\]

\[
= \frac{(-1)^k}{2^{k+1}k!} \sum_{m=0}^{\ell-1} \binom{\ell}{m} (-1)^m \int_{-\infty}^0 e^{z(m+3/2)} z^{k+1} \, dz
\]

\[
= \frac{(-1)^k}{2^{k+1}k!} \sum_{m=0}^{\ell-1} \binom{\ell}{m} (-1)^m \int_{-\infty}^0 e^{u} w^{k} \, dw
\]

\[
= \sum_{m=0}^{\ell-1} \binom{\ell}{m} (-1)^m \frac{1}{(2m+3)^{k+1}}.
\]

Notice that from (3.3.12),

\[
\sum_{\ell=1}^{\infty} p_{k+1,t} = \frac{(-1)^k}{2^{k+1}k!} \int_0^1 \frac{\log^k x}{\sqrt{x}} \, dx.
\]

With the substitution \( u = -(\log x)/2 \), we obtain

\[
\frac{1}{k!} \int_0^\infty e^{-u} u^k \, du = 1.
\]

Therefore, the length of the \( k \)th cycle formed in the process is bounded in probability, when \( k \) is fixed.
CHAPTER 4
THE BIPARTITE NEAREST NEIGHBOR GRAPHS

4.1 Introduction

Consider the complete bipartite graph $K_{n,n}$, on vertex sets $I = J = \{1, 2, \ldots, n\}$, in which each edge is assigned a length, or cost, $C_{i,j}$, which are iid continuous random variables. Color an edge green if it is one of the $k$ shortest adjacent to either end vertex, and blue otherwise. The graph made up of the green edges is called the bipartite $k^{\text{th}}$ nearest neighbor graph, denoted $B_k$. Another way of obtaining the graph $B_k$ is by randomly selecting, without replacement, edges of $K_{n,n}$ and coloring an edge green if it is one of the first $k$ edges selected incident to either end vertex, and blue otherwise.

Our work complements a study of the non-bipartite $k^{\text{th}}$ nearest neighbor graphs, denoted $O_k$, undertaken by Cooper and Frieze [8]. They show that $O_2$ is either connected or consists of a giant component and several small cycles with bounded total size whp, and that $O_3$ is connected whp. We obtain the same results for $B_2$ and $B_3$. 
Since we are working with bipartite graphs a study of the matching properties of these graphs is made accessible using Hall's Lemma, which states that there is a perfect matching in a bipartite graph iff every set of $s$ vertices from one part has at least $s$ neighbors in the other, for $1 \leq s \leq n$. We are able to show that $B_3$ has a perfect matching whp. We also show that $B_2$ does not have a perfect matching, but conjecture that the fraction of vertices matched goes to 1 whp, as suggested by computer experiments.

One reason that matchings in these bipartite graphs are so important comes from the following problem. Suppose there are $n$ workers available to fill $n$ jobs, where each worker is to be assigned to exactly one job. Also suppose that, for each worker $i$, we have a measurement of how qualified the individual is for each job $j$, given by $C_{i,j}$. An assignment of workers to jobs, then, is simply a permutation $\sigma$ of $\{1, 2, \ldots, n\}$. The cost of an assignment is given by

$$\sum_{i=1}^{n} C_{i,\sigma(i)}.$$ 

An optimal assignment is one with minimal cost.

We can think of the workers and jobs as vertices of a bipartite graph, with the $C_{i,j}$ as weights on the edges. Then the optimal assignment problem is equivalent to finding a perfect matching with minimal weight in this complete bipartite graph.

A particular assignment problem that has received a lot of attention is when the edges of $K_{n,n}$ are assigned weights $C_{i,j}$ where the $C_{i,j}$ are uniformly distributed $[0, 1]$ random variables. A conjecture of Parisi is that an optimal assignment has weight approaching $\pi^2/6$ as $n \to \infty$. (In fact, if the $C_{i,j}$ are exponentially distributed...
with parameter 1, he conjectures that the expected cost of an optimal assignment is exactly $1 + 1/2^2 + 1/3^2 + \cdots + 1/n^2$ for all $n$.) Lazarus [22] gives an asymptotic lower bound of $1 + 1/e$ for this problem. Walkup [42] gives an upper bound of 3, and Karp [18] improves the upper bound to 2.

One way of obtaining upper bounds is by finding matchings in reduced graphs (resulting in suboptimal matchings). Consider the following idea of Walkup [42]. Define random variables $Y_{i,j}$ and $Z_{i,j}$ such that $C_{i,j} = \min(Y_{i,j}, Z_{i,j})$ has a uniform $[0,1]$ distribution. The edges of $K_{n,n}$ now have two weights associated with them, and we imagine the $Y$'s visible to the workers and the $Z$'s visible to the jobs. Next, the complete bipartite graph is replaced by a new graph as follows. Each job and worker chooses the two edges of least weight (from their perspective) to which it is incident. The chosen edges make up the new graph. It can be shown that with high probability this graph has a perfect matching. Now, the average weight of the $Y$'s and $Z$'s in this reduced graph is about $3/n$, which leads to the upper bound. (Larger graphs are also looked at to cover the case when there is not a perfect matching. However, the probability that one needs to look at larger graphs and the weights of edges in the larger graphs are both sufficiently small so as not to ruin the bound of 3.) The actual average weight of the edges is less than $3$ whp, since each edge's weight is the minimum of a $Y$ and $Z$.

Notice that $B_k$ is also a reduced graph with short edges. Unlike in Walkup's setup, our graph is based on the $C_{i,j}$ directly. The analysis of $B_k$ is much more difficult, however, due to the dependence between edges. For example, if $\{i,j\}$ is
the shortest edge incident with vertex $i$ of part 1, then \{i, j\} is rather likely to be
one of the shortest edges incident with vertex $j$ of part 2. The conditioning device
is used to help handle this dependency (cf. Pittel [30]). It is hoped that by studying
these graphs we will better understand the random costs assignment problem, and
perhaps derive better bounds on the optimal assignment.

We will often refer to the $C_{i,j}$ as entries in an $n \times n$ matrix, in which case the
edges of the $k^{th}$ nearest neighbor graph correspond to the entries which are one of
the $k$ smallest in its row or in its column. We consider the case in which the $C_{i,j}$ are
iid exponential random variables with expectation $n$. The exponential was chosen
because it simplifies some of the computations; the actual structure of the graph
depends only on the order of the $C_{i,j}$ and would be the same for any iid continuous
random variable.

In section 2 we explore the forest $B_1$. We derive a formula for the expected
number of trees of each size and also show that the expected matching number is
approximately 0.8 whp.

In section 3 we study the graph $B_2$. We show that it is either connected or
consists of a large component and one or more cycles of a bounded total size whp.
Since the small components are even cycles, we may hope that the graph has a
perfect matching. We are able to show that it does not and that it actually has
at least $\frac{2\log n}{9\log \log n}$ vertices not matched whp. Finally we obtain a formula for the
probability that $B_2$ is connected.
In section 4 we show $B_3$ (and thus $B_k$ for $k > 3$) is connected and has a perfect matching whp.

4.2 Matching number of $B_1$

The graph $B_1$ is a forest consisting of $T$ trees, where $T$ is a random variable depending on the $C_{i,j}$. Consider the shortest edge in a particular tree. This edge connects a pair of vertices $i$ and $j$ with the property that $j$ is the nearest neighbor of $i$, and $i$ is the nearest neighbor of $j$. Every tree has exactly one pair of vertices with this property, and we will refer to the edge connecting them as the root edge of the tree.

**Proposition 4.2.1.** Let $\omega(n) \to \infty$. Then

$$T = \frac{n}{2} + O(\omega(n)\sqrt{n}),$$

or,

$$T = \frac{n}{2} + O_p(\sqrt{n}),$$

meaning that

$$\frac{T - n/2}{\sqrt{n}}$$

is bounded in probability.

**Proof.** Letting $I_{i,j} = 1$ if $C_{i,j}$ is the smallest entry in its row and column (i.e., if \{i, j\} is a root edge), and 0 otherwise, we get $T = \sum I_{i,j}$. We then calculate

$$E(T) = \sum E(I_{i,j}) = n^2 \frac{1}{2n-1} = \frac{n}{2} + O(1).$$
To calculate the variance of $T$, consider a pair of edges $e = \{i, j\}$ and $e' = \{i', j'\}$, $i \neq i'$, $j \neq j'$, with weights $u_1$ and $u_2$ respectively, and $u_1 < u_2$. The other $2n - 2$ entries in row $i'$ and column $j'$ must be larger than $u_2$, and the remaining $2n - 4$ entries in row $i$ and column $j$ must be larger than $u_1$. We obtain

$$E[T(T - 1)] = 2n^2(n - 1)^2 \int_0^\infty \int_0^\infty e^{-\frac{2n-4}{n}u_1}e^{-\frac{2n-2}{n}u_2}\frac{1}{n}e^{-\frac{u_1}{n}}e^{-\frac{u_2}{n}}du_2du_1$$

$$= 2(n - 1)^2 \int_0^\infty \int_0^\infty e^{-\frac{2n-3}{n}u_1}e^{-\frac{2n-1}{n}u_2}du_2du_1$$

$$= 2(n - 1)^2 \int_0^\infty \frac{n}{2n-1}e^{-\frac{4n-4}{n}u_1}du_1$$

$$= 2(n - 1)^2 \frac{n}{2n-1}4(n-1) = \frac{n^2(n-1)}{2(2n-1)}.$$

(4.2.1)

Thus,

$$E[T^2] = \frac{n^2(n - 1)}{2(2n-1)} + \frac{n^2}{2n-1} = \frac{n^2(n + 1)}{2(2n-1)}$$

and,

$$\text{Var}[T] = \frac{n^2(n + 1)}{2(2n-1)} - \left( \frac{n^2}{2n-1} \right)^2$$

$$= \frac{n^2(n - 1)}{2(2n-1)^2} = O(n). \quad \square$$

We conjecture that the number of trees actually converges to a normal distribution with mean $n/2$ and variance $n/8$. A formula for the $r$th factorial moment, arrived at by integrating as in (4.2.1), is given by

$$E[T(T - 1) \cdots (T - r + 1)] = \frac{n^2(n - 1)^2 \cdots (n - r + 1)^2}{(2n-1)(2n-2) \cdots (2n-r)}.$$ 

Computations of the first few central moments are in line with those of a normal distribution with given parameters.
Now let's investigate the size of the trees. Let $T_k$ be the number of trees with $k$ vertices. The smallest possible tree consists of two vertices; it is an isolated root edge. Let $A$ be the event that edge $\{1,1\}$ is isolated, and define

$$U = C_{1,1}$$
$$Y_i = \min_{j>1} (C_{i,j})$$
$$Z_j = \min_{i>1} (C_{i,j}).$$

(Notice that the $Y_i$ and $Z_j$ have an Exp($\lambda$) distribution with $\lambda = \frac{n-1}{n}$.) Then

$$E[T_2] = n^2 P(A),$$

and, since $A$ occurs iff every other entry in the first column (resp. row) is larger than the smallest of the other entries in its row (resp. column),

$$P(A) = E\left(\prod_{i=2}^{n} e^{-\frac{1}{n} \max(U,Y_i)} \prod_{j=2}^{n} e^{-\frac{1}{n} \max(U,Z_j)}\right)$$
$$= E\left(e^{-\frac{1}{n} (\sum_{i=2}^{n} \max(U,Y_i) + \sum_{j=2}^{n} \max(U,Z_j))}\right)$$
$$= E\left[E\left(e^{-\frac{1}{n} (\sum_{i=2}^{n} \max(u,Y_i) + \sum_{j=2}^{n} \max(u,Z_j))}\right) | U = u\right].$$

Now, if $X$ has an Exp($\lambda$) distribution,

$$E[\max(u,X)] = u P(X < u) + \int_{u}^{\infty} x \lambda e^{-\lambda x} \, dx$$
$$= u(1 - e^{-\lambda u}) + u e^{-\lambda u} + e^{-\lambda u} / \lambda$$
$$= u + e^{-\lambda u} / \lambda = u + e^{-u} + O(1/n).$$
Also,

$$E[\max(u, X)] = u^2 P(X < u) + \int_u^{\infty} x^2 e^{-\lambda x} dx$$

$$= u^2 (1 - e^{-\lambda u}) + u^2 e^{-\lambda u} + \frac{2u}{\lambda} e^{-\lambda u} + \frac{2}{\lambda^2} e^{-\lambda u}$$

$$= u^2 + \frac{2u}{\lambda} e^{-\lambda u} + \frac{2}{\lambda^2} e^{-\lambda u},$$

so that

$$\text{Var}[X] = u^2 + \frac{2u}{\lambda} e^{-\lambda u} + \frac{2}{\lambda^2} e^{-\lambda u} - \left( u + \frac{1}{\lambda} e^{-\lambda u} \right)^2$$

$$= \frac{2}{\lambda^2} e^{-\lambda u} - \frac{1}{\lambda^2} e^{-2\lambda u} = e^{-u} O(1).$$

Let $S_n = \sum_{i=2}^{n} \max(u, Y_i)$. Then, since the $Y_i$ are independent, Chebyshev's inequality yields

$$P(|S_n - E[S_n]| > n^{2/3}) \leq \frac{\text{Var}[S_n]}{n^{4/3}} = e^{-u} O(n^{-1/3}). \quad (4.2.2)$$

Then,

$$P(A) = E \left[ \exp \left( -2 \frac{n-1}{n} (U + e^{-U}) + O(n^{-1/3}) \right) + e^{-U} O(n^{-1/3}) \right]$$

$$= \int_0^\infty \left( \exp \left( -2 \frac{n-1}{n} (u + e^{-u}) \right) + e^{-u} O(n^{-1/3}) \right) \frac{1}{n} e^{-u} du$$

$$= \frac{1}{n} \int_0^\infty \left( e^{-2(u+e^{-u})+uO(1/n)} + e^{-u} O(n^{-1/3}) \right) du$$

$$= \frac{1}{n} \int_0^\infty \left( e^{-2(u+e^{-u})} + u e^{-u} O(1/n) + e^{-u} O(n^{-1/3}) \right) du$$

$$= \frac{1}{n} \int_0^\infty e^{-2(u+e^{-u})} du + O(n^{-4/3})$$

$$= \frac{1}{n} \int_0^1 x e^{-2x} dx + O(n^{-4/3})$$

$$= \frac{1 - 3e^{-2}}{4n} + O(n^{-4/3}).$$

53
And therefore,
\[ E[T_2] = (1 - 3e^{-2})\frac{n}{4} + O(n^{2/3}). \]

Now consider a tree with \( k \) vertices. If the edges are exposed from smallest to largest the root edge will be exposed first. The next edge exposed must be connected to one of the end vertices of the root edge, and so there are two places it may appear. In general, the \( m^{th} \) edge of the tree may be connected at any of the \( m \) vertices adjacent to previously exposed edges. A tree generated in this fashion is called a recursive tree. Notice that there are \((k - 1)! \) recursive trees on \( k \) vertices.

After the vertex of attachment is chosen for the next edge, there are \( n + O(k) \) vertices available to complete the edge. Let \( A_k^* \) be the event that on a particular set of \( k \) vertices there is an isolated tree isomorphic to a given recursive tree (where isomorphic includes edge appearance order). Then we write

\[ E[T_k^*] = (n + O(k))^k P(A_k^*) = n^k P(A_k^*)(1 + O(k^2/n)), \]

where \( T_k^* \) is the total number of such trees. It will soon be clear that the probability that there is a tree on a set of \( k \) vertices isomorphic to any particular recursive tree is almost constant, so that

\[ E[T_k] \sim (k - 1)! n^k P(A_k). \]

To see what restrictions the existence of a particular tree in \( B_1 \) puts on the weights of the other edges in the complete graph, let’s focus on a small example. Take \( n = 6 \), and suppose we want to calculate the probability of existence (in
$B_1$ of the (isolated) tree with edges exposed in the order $\{3,3\}, \{3,5\}, \{6,5\}$, with weights $u_1$, $u_2$ and $u_3$ respectively. The incidence matrix of the forest $B_1$ must look as follows, where present edges are represented by their weights, and $x$’s indicate potential edges which would be contained in other trees.

$$
\begin{pmatrix}
   x & x & 0 & x & 0 & x \\
   x & x & 0 & x & 0 & x \\
   0 & 0 & u_1 & 0 & u_2 & 0 \\
   x & x & 0 & x & 0 & x \\
   x & x & 0 & x & 0 & x \\
   0 & 0 & 0 & 0 & u_3 & 0 \\
\end{pmatrix}
$$

(4.2.4)

Notice that the root edge forces a row AND column of the incidence matrix to be zero while the addition of each new edge of the tree forces a new row OR column (but not both) to be zero (except for a few entries corresponding to edges in the tree).

If we define $Y_i$ and $Z_j$ similar to what was done before, we will be able to find an expression for the probability that the appropriate rows and columns of the incidence matrix are filled with zeros. In general, let

$$Y_i = \min_j (C_{i,j})$$

$$Z_j = \min_i (C_{i,j})$$

where the minimums are taken over all edges with neither end vertex belonging to the tree in question.

Suppose row $r$ is the row that must be filled with zeros because of the addition of an edge with weight $u_\ell$. Let $R$ be the event that each entry in the row is 0 (considering only those entries which are in columns corresponding to vertices not in the tree of interest).
Notice that (4.2.2) is true regardless of how many terms are missing from $S_n$ (here as many as $k$). So by examining the error terms in the calculation of $P(A)$ immediately following (4.2.2), if

$$f(x) = e^{-x} - e^{-2x},$$

and if $k = O(n^{2/3})$, then

$$P(R|u_\ell) = E \left[ e^{-\frac{1}{n} \sum_{j=1}^{\text{max}(u_\ell, z_j)}} \right] = f(u_\ell) + e^{-u_\ell} O(n^{-1/3}), \quad (4.2.5)$$

where the sum is over all vertices $j$ not in the tree. A similar calculation gives the same expression for the probability $P(C)$ that a column contains the appropriate zeros. Notice that the $R$'s and $C$'s are nearly independent, since the right side of (4.2.5) depends only on $u_\ell$. Now, we've specifically excluded considering entries where a row and column under consideration intersect. In many cases, the entry of intersection corresponds to an edge of the tree, so its weight is being conditioned on. But sometimes we also need a zero at the point of intersection (e.g., $a_{6,3} = 0$ in the matrix (4.2.4)). However, when conditioning on the $u_\ell$'s, such an entry will
contain a zero with probability at least $e^{-\max(u_t)/n}$. We obtain

$P(A_k) \sim \int_0^\infty \int_0^\infty \cdots \int_{u_{k-2}}^\infty f^2(u_1)f(u_2)f(u_3) \cdots$

$\cdots f(u_{k-1}) \frac{e^{-\frac{u_1}{n}}}{n} \frac{e^{-\frac{u_2}{n}}}{n} \cdots \frac{e^{-\frac{u_{k-1}}{n}}}{n} du_{k-1} \cdots du_2 du_1$

$\sim \frac{n^{-k+1}}{(k-2)!} \int_0^\infty \int_{u_1}^\infty \cdots \int_{u_{k-2}}^\infty \frac{f^2(u_1)}{n} \frac{f(u_2)}{n} \cdots \frac{f(u_{k-1})}{n} du_{k-1} \cdots du_2 du_1$

$= \frac{n^{-k+1}}{(k-2)!} \int_0^\infty f^2(x) \left( \int_x^\infty f(y)dy \right)^{k-2} dx$

$= \frac{n^{-k+1}}{(k-2)!} \int_0^\infty f^2(x) \left( \int_x^\infty e^{-y-x}dy \right)^{k-2} dx$

$= \frac{n^{-k+1}}{(k-2)!} \int_0^\infty e^{-2x-2e^{-x}} \left( 1 - e^{-e^{-x}} \right)^{k-2} dx,$

and with the substitution $u = e^{-e^{-x}},$

$= - \frac{n^{-k+1}}{(k-2)!} \int_{e^{-1}}^1 u \log u (1-u)^{k-2} du$

(4.2.6)

$= - \frac{n^{-k+1}}{(k-2)!} \int_{e^{-1}}^1 u \log u \sum_{i=0}^{k-2} \binom{k-2}{i} (-1)^i u^i du$

$= \frac{n^{-k+1}}{(k-2)!} \int_{e^{-1}}^1 \log u \sum_{i=1}^{k-1} \binom{k-2}{i-1} (-1)^i u^i du$

$= \frac{n^{-k+1}}{(k-2)!} \sum_{i=1}^{k-1} (-1)^i \binom{k-2}{i-1} \int_{e^{-1}}^1 u^i \log ud\alpha$

$= \frac{n^{-k+1}}{(k-2)!} \sum_{i=1}^{k-1} (-1)^i \binom{k-2}{i-1} \left[ \left( \frac{u^{i+1}\log u}{i+1} - \frac{u^{i+1}}{(i+1)^2} \right) \right]_{e^{-1}}^1$

$= \frac{n^{-k+1}}{(k-2)!} \sum_{i=1}^{k-1} (-1)^i \binom{k-2}{i-1} \frac{1-(i+2)e^{-(i+1)}}{(i+1)^2}$

Now using (4.2.3) we obtain,

$E[T_k] \sim n(k-1) \sum_{i=2}^{k} (-1)^i \binom{k-2}{i-2} \frac{1-(i+1)e^{-i}}{i^2}.$
Then, using the error term from (4.2.5), we have

$$E[T_k] = n(k - 1) \sum_{i=2}^{k} (-1)^i \binom{k - 2}{i - 2} \left( \frac{1 - (i + 1)e^{-i}}{i^2} \right) + O(kn^{2/3}).$$  \hspace{1cm} (4.2.7)

Some particular values are

- $E[T_2] \approx .1485n$
- $E[T_3] \approx .1190n$
- $E[T_4] \approx .0819n$
- $E[T_5] \approx .0540n$
- $E[T_6] \approx .0349n$.

If $p_k$ is the limiting probability that a root edge belongs to a tree with $k$ vertices, we can find the probability generating function for $p_k$ as follows. For a particular edge $e$ we simply multiply the number of trees that could be rooted on that edge by the probability of each, $P(A_k)$, and divide by the probability that the edge $e$ is actually a root edge.

$$p_k \sim \frac{(k - 1)!n^{k-2}P(A_k)}{1/(2n - 1)}$$

$$\sim 2(k - 1) \int_{e^{-1}}^{1} (u \log \frac{1}{u})(1 - u)^{k-2} du$$

58
so that,

\[
\sum_{k=2}^{\infty} p_k z^k = 2 \int_{e^{-1}}^{1} \left( u \log \frac{1}{u} \right) \frac{1}{u} \sum_{k=2}^{\infty} (k-1)(1-u)^{k-2} z^k \, du \\
= -2z \int_{e^{-1}}^{1} \left( u \log \frac{1}{u} \right) \frac{d}{du} \sum_{k=1}^{\infty} ((1-u)z)^{k-1} \, du \\
= -2z \int_{e^{-1}}^{1} \left( u \log \frac{1}{u} \right) \frac{1}{dt} \frac{1}{1-(1-u)z} \, du \\
= 2z^2 \int_{e^{-1}}^{1} \left( u \log \frac{1}{u} \right) \frac{1}{(1-(1-u)z)^2} \, du.
\]

Now, notice that the error term in (4.2.7) is dominant for large \( k \). However, we can show that whp there are no large trees in the graph. Notice, using a factor from (4.2.6), that the expected number of trees of size \( k \) may be bounded by

\[
n(k-1) \int_{e^{-1}}^{1} u \log u^{-1}(1-u)^{k-2} \, du \leq n(k-1) \int_{e^{-1}}^{1} (1-u)^{k-2} \, du \\
= n(1-e^{-1})^{k-1}.
\]

Then the expected number of vertices on trees with at least \( n^{1/12} \) vertices can be bounded by

\[
\sum_{k>n^{1/12}} nk(1-e^{-1})^{k-1} = O \left( n n^{1/12} (1-e^{-1})^{n^{1/12}} \right) \to 0.
\]

Therefore, whp, there are no vertices on trees of size at least \( n^{1/12} \).

Now we wish to compute the matching number of \( B_1 \). We do this by finding the matching number of the trees that make up the forest \( B_1 \). Notice that our computation of \( E[T_k] \), leading to (4.2.7), could have been carried out for any specific recursive tree. Then

\[
E[T_k^*] \approx \frac{E[T_k]}{(k-1)!}.
\]
To find the matching number of each recursive tree of each size is unreasonable. Instead, we use the fact that the expected number of each of the \((k - 1)!\) recursive trees of size \(k\) is about the same, and find the matching number of a randomly chosen recursive tree of size \(k\). According to our error term in (4.2.7), our expectations are off by a total of at most

\[
\sum_{k<n^{1/12}} O(kn^{2/3}) = O(n^{5/6}).
\]

Now, the number of vertices matched on a set of vertices of size \(O(n^{5/6})\) is \(O(n^{5/6})\), and so does not affect the limiting expected fraction of vertices matched.

Following Meir and Moon [20] we will find the expected node independence number of the random tree in question (here recursive trees), and then use the fact that \(n - N(T) = M(T)\) for any tree, where \(N(T)\) is the node independence number of the tree \(T\), and \(M(T)\) is the edge independence, or matching, number of the tree.

We will find a generating function that gives the number of recursive trees on \(k\) vertices that have a node independence number of \(t\). We split the number of these into two types. Type I trees are those for which every maximal set of independent vertices includes the root and type II trees are those for which at least one maximal set of independent vertices does not include the root. Let \(T\) denote a tree that is rooted at a vertex \(r\). If the root \(r\) is removed, we obtain a set of trees \(U_1, \ldots, U_j\) rooted at the vertices originally joined to \(r\). Notice that if each of \(U_1, \ldots, U_j\) is a type II tree, then \(T\) is type I and

\[
N(T) = 1 + \sum_{i=1}^{j} N(U_i), \quad (4.2.8)
\]
while if at least one of the trees \( U_1, \ldots, U_j \) is type I, then \( T \) is type II and

\[
N(T) = \sum_{i=1}^{j} N(U_i). \tag{4.2.9}
\]

Now we let \( y_{t,k} \) denote the number of rooted recursive trees \( T \) on \( k \) vertices with \( N(T) = t \), and let \( g_{t,k} \) and \( f_{t,k} \) be the number of these trees that are of types I and II respectively. Consider the following generating functions.

\[
G = G(z, x) = \sum_{t=1}^{\infty} \left( \sum_{k=1}^{t} g_{t,k} x^t \right) \frac{x^k}{k!}
\]

\[
F = F(z, x) = \sum_{t=1}^{\infty} \left( \sum_{k=1}^{t} f_{t,k} x^t \right) \frac{x^k}{k!}
\]

Since \( g_{t,k} + f_{t,k} = y_{t,k} \) and \( \sum_{t} y_{t,k} = (k-1)! \), the number of rooted recursive trees on \( k \) vertices, these series converge for \( |z| \) and \( |x| \leq 1 \) and

\[
Y = G + F.
\]

Now we can prove a number of relationships among these generating functions.

**Lemma 4.2.2.** \( G, F \) and \( Y \) satisfy

\[
\frac{dG}{dx} = ze^F, \tag{4.2.10}
\]

\[
\frac{dF}{dx} = (e^G - 1)e^F, \tag{4.2.11}
\]

\[
\frac{dY}{dx} = e^Y + (z - 1)e^F. \tag{4.2.12}
\]

**Proof.** To prove (4.2.10) we use (4.2.8) and sum over the different forests that may result when the root is removed. This involves summing over the number of trees
in the forest (they are all type II and their order is not important), over the ways in which the vertices may be divided among the trees and over the ways in which the matching number is attributable to each tree (which, since \( T \) is type I here, yields a total contribution of \( t - 1 \)). We then take the product, over the number of trees in the forest, of the number of type II trees satisfying these properties. We obtain

\[
g_{t,k} = \sum_{i \geq 0} \frac{1}{i!} \sum_{t_1 + \cdots + t_i = t-1 \atop k_1 + \cdots + k_i = k-1} \frac{(k-1)!}{k_1! \cdots k_i!} f_{t_1,k_1} \cdots f_{t_i,k_i}.
\]

Then we have

\[
d\frac{dG}{dx} = \sum_{k=1}^{\infty} \sum_{t=1}^{k} g_{t,k} z^t \frac{x^{k-1}}{(k-1)!}
\]

\[
= \sum_{k=1}^{\infty} \sum_{t=1}^{k} \sum_{i \geq 0} \frac{1}{i!} \sum_{t_1 + \cdots + t_i = t-1 \atop k_1 + \cdots + k_i = k-1} \frac{(k-1)!}{k_1! \cdots k_i!} f_{t_1,k_1} \cdots f_{t_i,k_i} z^t \frac{x^{k-1}}{(k-1)!}
\]

\[
= \sum_{i \geq 0} \frac{1}{i!} \sum_{k=1}^{\infty} \sum_{t=1}^{k} z^{t+i} \sum_{t_1 + \cdots + t_i = t-1 \atop k_1 + \cdots + k_i = k-1} \prod_{j=1}^{i} f_{t_j,k_j} z^{t_j} \frac{x^{k_j}}{k_j!}
\]

\[
= z \sum_{i \geq 0} \frac{1}{i!} \prod_{j=1}^{i} \left( \sum_{k_j=1}^{\infty} \sum_{t_j=1}^{k_j} f_{t_j,k_j} z^{t_j} \frac{x^{k_j}}{k_j!} \right)^{j}
\]

\[
= ze^{F}.
\]

To prove (4.2.11) we use (4.2.9) and sum over the different forests that may result, this time realizing that there is at least one tree of type I and possibly some type II trees. Notice that here the contribution to the matching number of the trees
in the forest is \( t \), since \( T \) is type II. We obtain

\[
f_{t,k} = \sum_{i \geq 1} \sum_{j \geq 0} \frac{1}{i! j!} \sum_{\substack{t_1 + \cdots + t_i + j = t \\ k_1 + \cdots + k_{i+j} = k-1}} (k-1)! \ell \cdot g_{t_1,k_1} \cdots g_{t_i,k_i} \cdot f_{t_{i+1},k_{i+1}} \cdots f_{t_{i+j},k_{i+j}}.
\]

Then we have

\[
\frac{dF}{dx} = (e^G - 1)e^F.
\]

Finally,

\[
\frac{dY}{dx} = \frac{d}{dx}(G + F) = ze^F + (e^G - 1)e^F = ze^F + e^{(G+F)} - e^F = e^Y + (z - 1)e^F. \quad \Box
\]

Now we set \( y = y(x) = Y(1,x) \), \( g = g(x) = G(1,x) \) and \( f = f(x) = F(1,x) \).

Then, from Lemma 4.2.2, we have

\[
\frac{dg}{dx} = e^f
\]

\[
\frac{df}{dx} = (e^g - 1)e^f = e^y - e^f
\]

\[
\frac{dy}{dx} = e^y.
\]

These equations can be solved, using \( g(0) = f(0) = y(0) = 0 \), and we obtain

\[
g(x) = \log(1 - \log(1 - x)),
\]

\[
f(x) = -\log((1 - x)(1 - \log(1 - x))), \quad \text{(4.2.13)}
\]

\[
y(x) = -\log(1 - x).
\]
Now let $\mu(k)$ be the expected vertex independence number of a random recursive tree on $k$ vertices. Then

$$\mu(k) = \frac{\sum_{t=1}^{k} \ell y_t k}{(k-1)!}. $$

Now define

$$M(x) = \sum_{k=1}^{\infty} \mu(k)(k-1)! \frac{x^k}{k!} = \left( \frac{dY}{dz} \right)_{z=1}. \quad (4.2.14)$$

Now,

$$\frac{d}{dx} M(x) = \frac{d}{dx} \frac{dY}{dz} \bigg|_{z=1} = \left( \frac{d}{dz} \frac{dY}{dx} \right)_{z=1} = \frac{d}{dz} \bigg|_{z=1} \left( e^Y + (z-1)e^F \right) = \left( e^Y \frac{dY}{dz} + e^F + (z-1)e^F \frac{dF}{dz} \right)_{z=1} = e^Y M(x) + e^f.$$ 

Then using (4.2.13) we obtain

$$M'(x) = \frac{1}{1-x} M(x) + \frac{1}{(1-x)(1-\log(1-x))},$$

which has solution

$$M(x) = \frac{1}{1-x} \int_0^x \frac{1}{1-\log(1-y)} dy.$$ 

From (4.2.14),

$$\mu(k) = \frac{1}{(k-1)!} \frac{d^k}{dx^k} M(x).$$

Then from (4.2.7) and since the expected matching number is $k - \mu(k)$, we have the following.
Theorem 4.2.3. The expected matching number of \( B_1 \) is asymptotic to \( cn \), where \( c \) is given by

\[
\sum_{k=2}^{\infty} \left( (k-1) \sum_{i=2}^{k} (-1)^i \binom{k-2}{i-2} \left( \frac{1 - (i+1)e^{-i}}{i^2} \right) \cdot \left( k - \frac{1}{(k-1)!} \frac{d^k}{dx^k} \left( \frac{1}{1-x} \int_{0}^{x} \frac{1}{1 - \log(1-y)} dy \right) \right) \right].
\]

Numerically, \( c \approx 0.807 \).

4.3 Properties of \( B_2 \)

Theorem 4.3.1. With high probability, the graph \( B_2 \) is connected or consists of a giant component and small cyclic components with a bounded total size.

Proof. First a definition. An \((i, j)\) configuration, defined for \( i > j \), is a set of \( i \) rows and \( j \) columns where the two best choices of each of the \( i \) rows are among the set of \( j \) columns and also the best two choices of each of the other \( n - i \) rows are among the remaining \( n - j \) columns. A minimal configuration is minimal with respect to \( j \). Notice that every component contains at least one minimal configuration.

An important property of minimal configurations is that each of the \( j \) columns is selected at least once. Let \( M_{i,j} \) be the number of minimal \((i, j)\) configurations and \( E_{i,j} = E[M_{i,j}] \). We begin by showing that for \( j \) small there are no minimal configurations whp, or more precisely, we show that:

\[
\sum_{j=2}^{\left\lfloor \epsilon n \right\rfloor} \sum_{i=j+1}^{n} E_{i,j} + \sum_{j=\omega(n)}^{\left\lfloor \epsilon n \right\rfloor} E_{j,j} \to 0, \tag{4.3.1}
\]

where \( \epsilon \) may be chosen as small as we like, and \( \omega(n) \to \infty \) arbitrarily slowly.
We can bound a generic $E_{i,j}$ as follows.

$$E_{i,j} \leq \binom{n}{i} \binom{n}{j} \left( \frac{j}{n} \right)^{2i} \left( 1 - \frac{j}{n} \right)^{2(n-i)} c \sqrt{j} \left( 1 - e^{-\frac{2i}{j}} \right)^j$$  \hspace{1cm} (4.3.2)

$$\leq c \frac{n^{i+j} e^{i+j} j^{2i}}{i^{i+j} j^{i+j} n^{2i}} e^{-2i} e^{\frac{2ii}{n}} \sqrt{j} \left( 1 - e^{-\frac{2i}{j}} \right)^j$$

$$\leq c \left( \frac{e j}{n} \right)^{i-j} e^{\frac{2ii}{n}} \left( 1 - e^{-\frac{2i}{j}} \right)^j, \hspace{1cm} (4.3.3)$$

where $c$ is some absolute constant. Let's explain the source of the factors in (4.3.2).

For our calculations, we let each row chose a column for its smallest entry and then choose another column (with replacement!) for its second smallest entry. Clearly the expected number of components does not decrease in this scheme, since we see fewer edges than are actually present in the graph. Then all the factors are self-explanatory except perhaps the last. This factor appears because each of the first $j$ columns must be chosen by at least one of the first $i$ rows. We use the Poissonization technique. Notice that we are allocating $2i$ balls (choices of rows) into $j$ boxes (columns). We can think of the number of balls in each box as a Poisson r.v. $Z_k$, $1 \leq k \leq j$, with parameter $x$. Conditioned on the sum of these values equaling $2i$, the distribution of the balls is the same as if they were allocated at random. Then the probability that each box receives at least one ball is bounded by

$$\frac{\prod_{k=1}^j \mathbb{P}(Z_k > 0)}{\mathbb{P}(Z_1 + \cdots + Z_j = 2i)} = \frac{(1 - e^{-x})^j}{e^{-jx}(jx)^{2i}/(2i)!}.$$  \hspace{1cm} (4.3.4)

Letting $x = 2i/j$ we obtain the last factor in the expression, for some constant $c$.  

66
Beginning with the single sum in (4.3.1), we compute

\[ \sum_{j=[\omega(n)]}^{[en]} E_{j,j} \leq \sum_{j=[\omega(n)]}^{[en]} c e^{2i^2/n} (1 - e^{-2})^j \]

\[ \leq \sum_{j=[\omega(n)]}^{[en]} c e^{2\epsilon j} (1 - e^{-2})^j \]

\[ = C_\epsilon \left( e^{2\epsilon (1 - e^{-2})} \right)^{[\omega(n)]}, \]

which goes to zero, for sufficiently small \( \epsilon \), since \( \omega(n) \to \infty \). Here and in the rest of the proof \( C_\epsilon \) is a generic constant depending only on \( \epsilon \).

To show that the double sum in (4.3.1) goes to zero, we let \( k = i - j \), write

\[ \sum_{j=2}^{[en]} \sum_{k=1}^{n-j} E_{i,j} = \sum_{1}^{\ell} + \sum_{2}^{1} + \sum_{3}^{1} \]

and use (4.3.3) to bound the terms in each sum. To simplify the expressions, let \( \ell = \lfloor \log \log n \rfloor \). Note that in summing the geometric series involving \( k \) in each of the following expressions, we have made the assumption that \( \epsilon \) is chosen so that \( \frac{\epsilon j}{n} < \frac{1}{2} \) for all \( j \), i.e., \( \epsilon < \frac{1}{2\ell} \).

\[ \sum_{j=2}^{\ell} \sum_{k=1}^{n-j} E_{i,j} \leq \sum_{j=2}^{\ell} \sum_{k=1}^{n-j} c \left( \frac{\epsilon j}{n} \right)^k e^{2(i+k)j/n} \]

\[ \leq \sum_{j=2}^{\ell} 2c \left( \frac{\epsilon j}{n} \right) e^{2j} \leq \frac{2c\ell^2}{n} e^{2\ell} \to 0. \]
\[ \sum_{2} = \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} \sum_{k=j+1}^{n-j} E_{i,j} \]

\[ \leq \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} \sum_{k=j+1}^{n-j} c \left( \frac{e(j)}{n} \right)^k e^{2(j+k)j} \left( 1 - e^{-\frac{2(j+k)j}{n}} \right)^j \]

\[ \leq \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} \sum_{k=j+1}^{n-j} c \left( \frac{e(j)}{n} \right)^k e^{\frac{4j^2}{n}} (1 - e^{-4})^j \]

\[ \leq \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} c e^{\frac{4j^2}{n}} (1 - e^{-4})^j \]

\[ \leq \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} c (e^{4e}(1 - e^{-4}))^j \]

\[ = C_e (e^{4e}(1 - e^{-4}))^n \to 0 \]

for sufficiently small \( \epsilon \).

\[ \sum_{3} = \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} \sum_{k=j+1}^{n-j} E_{i,j} \]

\[ \leq \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} \sum_{k=j+1}^{n-j} c \left( \frac{e(j)}{n} \right)^k e^{\frac{2(j+k)j}{n}} \]

\[ \leq \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} \sum_{k=j+1}^{n-j} c \left( \frac{e(j)}{n} \right)^k e^{2j} \]

\[ \leq \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} c \left( \frac{e(j)}{n} \right)^j e^{2j} \]

\[ \leq \sum_{j=t+1}^{\lfloor \frac{en}{2} \rfloor} c (e^{3j})^j \to 0 \]

for sufficiently small \( \epsilon \).

Now we will show that for \( i \) and \( j \) of order \( n \) the expected number of minimal configurations is exponentially small. Let \( a = \frac{i}{n} \) and \( b = \frac{j}{n} \). Recall that \( a \geq b \).
Also note that we need only consider \( b \leq 0.5 \). Let's begin by rewriting (4.3.2) in terms of \( a \) and \( b \).

\[
E_{a,b} = \left( \frac{b^{2a}(1-b)^{2(1-a)}}{a^a(1-a)^{(1-a)b^b(1-b)}(1-e^{-2a})^b + o(1)} \right)^n
\]

We will show that \( F(a, b) \) attains a maximum value of 1 over the values of \( a \) and \( b \) considered. Taking logs of factors involving \( a \) we write

\[
f(a) = -a \log a - (1-a) \log (1-a) + 2a \log b + 2(1-a) \log (1-b),
\]

\[
f'(a) = -\log a + \log (1-a) + 2 \log b - 2 \log (1-b),
\]

\[
f''(a) = -\frac{1}{a} - \frac{1}{(1-a)}.
\]

Setting \( f'(a) = 0 \) and solving we find a unique maximum in \((0,1)\) of

\[
a = \frac{b^2}{(1-b)^2 + b^2} = \frac{b}{b^2 + b^2} = bg(b).
\]

Since \( a \geq b \), if \( g(b) \leq 1 \) we must take \( a = b \) to maximize \( f(a) \). Note \( g(b) \leq 1 \) iff \( 2b^2 - 3b + 1 \geq 0 \). The zeros of this upward opening quadratic equation are 0.5 and 1, and so we have that for \( b \) in \((0,0.5]\), \( g(b) \leq 1 \). Therefore, \( f(a) \) is maximized at \( a = b \). Observe that \( F(b,b) = 1 \).

We obtain

\[
E_{a,b} \leq \left( (1-e^{-2b})^b + o(1) \right)^n \to 0.
\]

To complete the proof, we need to show that the small components actually are cycles. Thus far we’ve only considered choices made by the rows. Suppose a set
of $i$ rows only choose entries among a set of $i$ columns. Condition on the values in these entries. Notice that the remaining entries in the intersection of these sets of rows and columns are variables which stochastically dominate the values in the other $n - i$ entries of the $i$ columns. Thus if one of the columns chooses an entry other than an entry chosen by one of the $i$ rows, the probability that it is among the $i$ rows is less than $i/n$. The expected number of sets of $i$ rows and $i$ columns in which the $i$ rows select entries only among the $i$ columns and at least one of the columns chooses an entry in the $i$ rows not chosen by any of the rows is bounded by

$$\binom{n}{i} \binom{i}{i} \left(\frac{i}{n}\right)^{2i} \frac{2i}{n},$$

since there are $i$ columns to choose from and each column makes two choices. We can further bound this expression and write

$$\frac{2i^{2i+2}}{n} \to 0,$$

if $i < \log \log n$.

Thus, if each column chooses entries among the $i$ rows, the rows and columns choose the same set of entries whp. Thus, whp, they form an isolated subgraph of degree 2, i.e., a set of cycles. □

We have shown that all the small components, if there are any, are even cycles. Thus all these vertices may be matched. Computer experiments suggest that most of the vertices may be matched, and it seems plausible that whp all the vertices may be matched. However, we can prove the following.
Theorem 4.3.2. With high probability, $B_2$ does not have a perfect matching. Moreover, in a maximal matching there are at least \( \frac{2 \log n}{13 \log \log n} \) vertices not matched with high probability.

Proof. To prove the theorem we will show that whp $B_2$ contains a specific configuration consisting of a set of $k$ vertices which has at most $m = k - c + 1$ neighbors, where $c = \left\lfloor \frac{\log n}{13 \log \log n} \right\rfloor$. Let $k = c \cdot \lfloor 20 \log n \rfloor$. Our configuration will consist of $k$ row and $m$ column vertices which form $c$ cycles of length $t = k/c$ joined at a single column vertex when considering only choices made by row vertices, and with no edges between the $k$ row vertices and remaining $n - m$ column vertices when considering choices made by both row and column vertices. Notice that if such a configuration exists, then there are at least $c - 1$ row (and thus column) vertices not matched in a maximal matching.

The first step is to accurately compute the expected number of these configurations. Let $X_i$ be the second smallest entry among the first $m$ entries of row $i$, $1 \leq i \leq k$, and $Y_j$ the second smallest among the last $n - k$ entries of column $j$, $m + 1 \leq j \leq n$. Letting $P_{nkm}$ be the probability that none of the first $k$ row vertices are connected to any of the last $n - m$ column vertices, we can write

\[
P_{nkm} = E \left( \prod_{i=1}^{k} e^{-\frac{1}{n} \sum_{j=m+1}^{n} \max(X_i,Y_j)} \right) = E \left( \prod_{i=1}^{k} E \left( e^{-\frac{1}{n} \sum_{j=m+1}^{n} \max(X_i,Y_j)} | X_i \right) \right). \tag{4.3.4}
\]

Now let $x = X_i$, $Z_j = \max(x, Y_j)$ and $S_n = \sum_{j=m+1}^{n} Z_j$. Notice that $Y_j$ is the second smallest of $n - k$ exponential ($\frac{1}{n}$) random variables, and so is asymptotically
distributed like the sum of two exponential (λ) random variables, where \( \lambda = \frac{n-k}{n} \).

The density and cumulative distribution functions are given by

\[
\begin{align*}
f_Y(y) &= \lambda^2 y e^{-\lambda y}, \\
F_Y(y) &= 1 - e^{-\lambda y} - \lambda y e^{-\lambda y}.
\end{align*}
\]  

(4.3.5)

We compute

\[
\begin{align*}
E[Z] &= x P(Y < x) + \int_x^\infty \lambda^2 y^2 e^{-\lambda y} dy \\
&= x (1 - e^{-\lambda x} - \lambda x e^{-\lambda x}) + \frac{1}{\lambda} \int_{\lambda x}^\infty y e^{-y} dy \\
&= x (1 - e^{-\lambda x} - \lambda x e^{-\lambda x}) + \frac{1}{\lambda} (-2 e^{-y} - 2 ye^{-y} - y^2 e^{-y}) \bigg|_{\lambda x}^\infty \\
&= x - xe^{-\lambda x} - \lambda x^2 e^{-\lambda x} + \frac{2}{\lambda} e^{-\lambda x} + 2xe^{-\lambda x} + \lambda x^2 e^{-\lambda x} \\
&= x + xe^{-\lambda x} + \frac{2}{\lambda} e^{-\lambda x},
\end{align*}
\]  

(4.3.6)

and so

\[E[S_n] = (n-m)(x + xe^{-\lambda x} + \frac{2}{\lambda} e^{-\lambda x}).\]

To compute the inner expectation of (4.3.4) we show that \( S_n \) is highly concentrated around its mean. Using Chebyshev's inequality, we have

\[
\begin{align*}
e^{\theta b(n-k)} P(S_n > b(n-m)) &\leq \frac{E[e^{\theta S_n}]}{\gamma_2} = (E[e^{\theta Z}])^{n-m}.
\end{align*}
\]  

(4.3.7)
Now, if $\theta < \lambda,$

$$E[e^{\theta Z}] = e^{\theta x}P(Y < x) + \int_x^\infty e^{\theta y} \lambda^2 ye^{-\lambda y} dy$$

$$= e^{\theta x}(1 - e^{-\lambda x} - \lambda x e^{-\lambda x}) + \frac{\lambda^2}{(\lambda - \theta)^2} \int_x^\infty ye^{-y} dy$$

$$= e^{\theta x}(1 - e^{-\lambda x} - \lambda x e^{-\lambda x}) + \frac{\lambda^2}{(\lambda - \theta)^2} (e^{-(\lambda - \theta)x} - (\lambda - \theta) xe^{-(\lambda - \theta)x})$$

$$= e^{\theta x}(\lambda x - 1 - \lambda x + \frac{\lambda^2}{(\lambda - \theta)^2} + \frac{\lambda^2 x}{(\lambda - \theta)})$$

$$= e^{\theta x} \left( 1 + \left( \frac{\lambda^2}{(\lambda - \theta)^2} - 1 + \frac{\lambda^2 x}{(\lambda - \theta)} \right) e^{-\lambda x} \right)$$

$$= e^{\theta x} \left( 1 + \left( \frac{2\lambda - \theta}{\lambda - \theta} + \frac{\lambda x}{\lambda - \theta} \right) \theta e^{-\lambda x} \right)$$

Letting $b = (1 + \epsilon)E[Z]$ in (4.3.7), we have

$$P(S_n > (1 + \epsilon)E[S_n]) \leq \exp \left[ -(n - m) \left( (1 + \epsilon)E[Z]\theta - \log(E[e^{\theta Z}]) \right) \right].$$

Now we bound the exponent making use of $\log(1 + y) \leq y$:

$$(1 + \epsilon)E[Z]\theta - \log(E[e^{\theta Z}])$$

$$= (1 + \epsilon)\theta (x + xe^{-\lambda x} + \frac{2}{\lambda} e^{-\lambda x}) - \theta x - \log \left( 1 + \left( \frac{2\lambda - \theta}{(\lambda - \theta)^2} + \frac{\lambda x}{\lambda - \theta} \right) \theta e^{-\lambda x} \right)$$

$$\geq (1 + \epsilon)\theta (x + xe^{-\lambda x} + \frac{2}{\lambda} e^{-\lambda x}) - \theta x - \left( \frac{2\lambda - \theta}{(\lambda - \theta)^2} + \frac{\lambda x}{\lambda - \theta} \right) \theta e^{-\lambda x}$$

$$= \theta \left( \epsilon \left( x + xe^{-\lambda x} + \frac{2}{\lambda} e^{-\lambda x} \right) + e^{-\lambda x} \left( x + \frac{2}{\lambda} - \frac{2\lambda - \theta}{(\lambda - \theta)^2} \right) \left( \frac{\lambda x}{\lambda - \theta} \right) \right)$$

$$= \theta \left( \epsilon \left( x + xe^{-\lambda x} + \frac{2}{\lambda} e^{-\lambda x} \right) - \theta e^{-\lambda x} \left( \frac{3\lambda - 2\theta}{\lambda(\lambda - \theta)^2} + \frac{x}{\lambda - \theta} \right) \right).$$

Recall that $\lambda = 1 + O(\frac{\log n}{n})$ and notice that if $\theta = o(\epsilon),$ then this expression is positive and $\Omega(\epsilon \theta).$ Then we have

$$P(S_n > (1 + \epsilon)E[S_n]) = O(e^{-n\epsilon \theta}).$$

(4.3.8)
A similar computation yields the same bound for the probability that $S_n$ is far below its expectation. From Chebyshev's inequality,

$$e^{-\theta b(n-k)}P(S_n < b(n-m)) \leq E(e^{-\theta S_n}) = (Ee^{-\theta Z})^{n-m}$$

so letting $b = (1 - \epsilon)E[Z]$ yields

$$P(S_n < (1 + \epsilon)E[S_n]) \leq \exp \left[ -(n - m) \left( -(1 - \epsilon)E[Z] - \log(E[e^{\theta Z}]) \right) \right].$$

Now bounding the exponent using $\log(1 - y) = -y + O(y^2)$,

$$-(1 - \epsilon)E[Z] - \log(E[e^{\theta Z}]) = \theta \left( \epsilon \left( x + xe^{-\lambda x} + \frac{2}{\lambda} e^{-\lambda x} \right) - \theta e^{-\lambda x} \left( \frac{3\lambda + 2\theta}{\lambda(\lambda + \theta)^2} + \frac{x}{\lambda + \theta} \right) + O(\theta) \right)$$

thus,

$$P(S_n < (1 - \epsilon)E[S_n]) = O(e^{-n\epsilon \theta}). \quad (4.3.9)$$

Now we rewrite the inner expectation of (4.3.4) as follows:

$$E[e^{-\frac{S_n}{n}}] = E \left[ e^{-\frac{S_n}{n}} I_{(S_n - E[S_n] < \epsilon E[S_n])} \right] + E \left[ e^{-\frac{S_n}{n}} I_{(S_n - E[S_n] > \epsilon E[S_n])} \right]$$

$$= \exp \left( -\frac{E[S_n]}{n} + O \left( \frac{E[S_n]}{n} \right) \right) \left( 1 + O(e^{-n\epsilon \theta}) \right) + O(e^{-n\epsilon \theta})$$

$$= \exp \left( -\left( 1 - \frac{k}{n} \right) \left( x + xe^{-\lambda x} + \frac{2}{\lambda} e^{-\lambda x} \right) + O(\epsilon x) \right) + O(e^{-n\epsilon \theta}).$$

With $\lambda = 1 + O(\log^2 n)$ and taking $\epsilon = n^{-1/4}$ and $\theta = n^{-1/2}$ we write

$$E[e^{-\frac{S_n}{n}}] = e^{-\left( x + xe^{-x} + 2e^{-x} \right) + O(xn^{-1/4})} + O(e^{-n^{1/4}}).$$
Now,

\[ P_{nkm} = E \left( \prod_{i=1}^{k} (e^{-(X_i + X_i e^{-X_i} + 2e^{-X_i}) + O(X_n n^{-1/4})} + O(e^{-n^{1/4}})) \right) \]

\[ = \left( E \left( e^{-(X + X e^{-X} + 2e^{-X}) + O(X_n n^{-1/4})} + O(e^{-n^{1/4}}) \right) \right)^k \]

where \( X \) is asymptotically distributed like the sum of two exponential(\( \mu \)) random variables, where \( \mu = \frac{m}{n} \). The density and cumulative distribution functions are as in (4.3.5) with \( \mu \) in place of \( \lambda \).

Now,

\[ E \left( e^{-(X + X e^{-X} + 2e^{-X}) + O(X_n n^{-1/4})} \right) = \int_0^\infty e^{-(x+xe^{-x}+2e^{-x})+O(xn^{-1/4})} \mu^2 xe^{-\mu x} dx \]

\[ = \mu^2 \int_0^{n^{1/8}} xe^{-(x+xe^{-x}+2e^{-x})+O(xn^{-1/4})} dx \]

\[ + \mu^2 \int_{n^{1/8}}^{\infty} xe^{-(x+xe^{-x}+2e^{-x})+O(xn^{-1/4})} dx \]

\[ = \mu^2 \int_0^{n^{1/8}} xe^{-(x+xe^{-x}+2e^{-x})} (1 + O(n^{-1/8})) dx \]

\[ + \mu^2 O \left( \int_{n^{1/8}}^{\infty} xe^{-x/2} dx \right) \]

\[ = \mu^2 (\rho(1 + O(n^{-1/8})) + O(n^{1/8} e^{-n^{1/8}})) \]

\[ = \mu^2 (\rho + O(n^{-1/8})) , \]

where

\[ \rho = \int_0^\infty xe^{-(x+xe^{-x}+2e^{-x})} dx \approx 0.530527 . \]

Therefore,

\[ P_{nkm} = \mu^{2k} (\rho + O(n^{-1/8}))^k = \left( \frac{m}{n} \right)^{2k} (\rho^k + O(n^{-1/9})) . \]

75
If \( T \) is the number of configurations now we can calculate \( E[T] \) as follows. From the \( n \) row vertices, we must select \( c \) sets of \( t \) vertices, one for each cycle in the configuration. From the \( n \) column vertices we choose \( t \) vertices for the first cycle and then \( t - 1 \) for each of the remaining \( c - 1 \) cycles, and we multiply by \( t \) for the number of positions on the first cycle to which the remaining ones might be attached. We then divide by \( c! \) since the order of the cycles doesn't matter. The number of ways to form each cycle is given by \( \frac{(t-1)^{\frac{c}{2}}}{2} \). Then we divide by the number of ways the \( k \) rows could have chosen their neighbors, \( \binom{m}{2} \), and finally multiply by \( P_{nkm} \). We obtain

\[
E[T] = \frac{n!}{(t!)^c(n-k)!} \cdot \frac{n!}{t!!((t-1)!)^c-1(n-m)!} \cdot \frac{t}{c!} \left( \frac{(t-1)!}{2} \right)^c \binom{m}{2}^{-k} P_{nkm}
\]

\[
\sim \frac{n^{k+m}}{c!2^c} \left( \frac{m}{2} \right)^{-k} \left( \frac{m}{n} \right)^{2k} \rho^k \sim \frac{(2\rho)^k}{c!2^c n^{c-1}}.
\]

Now,

\[
E[T] > \frac{(2\rho)^k}{c^n n^c}
\]

so that

\[
\log E[T] > k \log(2\rho) - c \log c - c \log n
\]

\[
\sim c(20 \log(2\rho) \log n - \log c - \log n) \to \infty.
\]

Now we will show that \( E[T^2] \sim E^2[T] \). To do this, we write \( E[T(T-1)] = E_1 + E_2 \), where \( E_1 \) is the contribution due to disjoint configurations and \( E_2 \) the contribution due to overlapping ones. First we compute the expected number of pairs of disjoint
configurations. We will in fact bound this expectation from above by requiring only that each of the $2k$ row vertices involved are not connected to any of the $n - 2m$ column vertices involved, and ignoring that each set of $k$ row vertices should also not be connected to any of the $m$ column vertices of the other configuration. We first choose vertices for the first configuration and form it. Then from the remaining $n - k$ row and $n - m$ column vertices a second configuration is formed. Then we divide by the number of ways the row vertices could have chosen neighbors from the $2m$ column vertices and multiply by the probability that none of the $2k$ row vertices are connected to the remaining column vertices, $P_{n,2k,2m}$. Then we have

$$E_1 \leq \frac{n!}{(t!)^c(n-k)!} \cdot \frac{n!}{t!(t-1)!c^{-1}(n-m)!} \cdot \frac{t}{c!} \left( \frac{(t-1)!t!}{2} \right)^c \cdot \left( \frac{2m}{2} \right)^{-2k} \cdot P_{n,2k,2m} \sim \frac{n^{2(k+m)}}{(c!)^22^{2c}} \left( \frac{2m}{2} \right)^{-2k} \left( \frac{2m}{n} \right)^{4k} \rho^{2k} \sim \frac{(2\rho)^{2k}}{(c!)^22^{2c}\rho^{2c-1}} \sim \mathbb{E}^2[T].$$

Now let's find an upper bound for the case in which the two configurations overlap. Let $c'$ be the number of cycles in the second configuration not entirely contained in the first. First we choose a set of $c - c'$ cycles of the first configuration to belong to the second, which can be done in at most $2^c$ ways. Then after choosing vertices for the first configuration we must choose additional vertices for each of these $c'$ cycle portions. We will complete the structure of the overlapping configurations by starting with the first configuration and sequentially appending cycles of the second.
Since the two configurations overlap, at least one of the cycles in the new configuration shares vertices with the first configuration, we append such a cycle first. Let \( v_1, v_2, \ldots, v_{c'} \) be the number of row vertices in each cycle not belonging to the first configuration, and \( v \) the total number of new row vertices. The number of column vertices to be chosen depends on the number of breaks in the cycle, that is, the number of times two nearest new row vertices are not connected to a new column vertex but are instead joined to vertices of the first configuration, or to the vertex shared by all cycles in the second configuration. Let \( b_1, b_2, \ldots, b_{c'} \) be the number of breaks in each of the cycles, and \( b \) the total number of breaks. Then \( v_i - b_i \) column vertices must be chosen for the \( i^{th} \) cycle of the second configuration. Now we must order the sets of vertices for each of the new cycles and choose places for the breaks in the new cycles, the latter of which may be done in \( {v_i - 1 \choose b_i} \) ways and may be bounded by \( t^{b_i} \). The first cycle can be appended in at most \( (k - c + 1)^{2b_1} \) ways, since at each break the two end vertices must be connected to one of the column vertices of the first configuration. Now we choose one of the \( t \) column vertices of this cycle at which each of the remaining cycles is to be connected. For the remaining cycles there are then at most \( (k - c + 2) \) places each end vertex can be attached, yielding an overall bound of \( t(k - c + 2)^{2b} \) on the number of ways to append the cycles of the second configuration. Again we divide by the number of ways the row vertices could have chosen their neighbors and then multiple by \( P_{n, k+u, k+u-c+1-b} \), and note that this is an upper bound that ignores the fact that there should also be no edges between the non-overlapping vertices of the two configurations, when
column choices are also considered. Finally, we sum over the choices for the values of the $v_i$'s and $b_i$'s, denoted by the * in the expression below. Then,

$$E_2 \leq \sum \sum_{c'}^{c} 2^c \frac{n!}{(t!)^c v_1! \ldots v_{c'}!(n-k-v)^c} \cdot \frac{n!}{(t-1)!^{c-1} (v_1-b_1)! \ldots (v_{c'}-b_{c'})!(n-k-v+c-1+b)!}$$

$$\cdot t \left( \frac{(t-1)!t!}{2} \right)^c v_1! \ldots v_{c'}!(v_1-b_1)! \ldots (v_{c'}-b_{c'})! t^{b_1} \ldots t^{b_{c'}}$$

$$\cdot t(k-c+2)^{2b} \left( \frac{k+v-c+1-b}{2} \right)^{-k+u} \frac{P_{n,k,v,k+v-c+1-b}}{2^{k+v}(k-v-c+1-b)^{2(k+v)} n^{2(k+v)}} \cdot \rho^{k+v}$$

$$\leq \sum \sum_{c'}^{c} n^{2(k+v)-c+1-b} \cdot \frac{t^{b+1}t^{2b}}{c!}$$

$$\cdot \frac{2^{k+v}}{(k+v-c+1-b)^{2(k+v)}} \cdot \frac{(k+v-c+1-b)^{2(k+v)}}{n^{2(k+v)}} \cdot \rho^{k+v}$$

$$\leq \sum \sum_{c'}^{c} \frac{t^{b+1}k^{2b}(2\rho)^{k+v}}{c!n^{c+b-1}}.$$ 

Now we make use of the fact that $b \geq c'$, since each cycle appended has at least one break, and that $v \leq tc'$, since each cycle appended has less than $t$ new row vertices, to bound this last expression by

$$\sum \sum_{c'}^{c} \frac{t^{c+1}k^{2c'}(2\rho)^{k+c'}}{c!n^{c+c'-1}}.$$ 

Notice that this bound is valid for any combination of the at most $k$ choices for each of the $v_i$ and $b_i$, a total of at most $2c'$ variables. A new bound is thus

$$\sum_{c'}^{c} \frac{t^{c+1}k^{4c'}(2\rho)^{k+c'}}{c!n^{c+c'-1}} = \frac{t(2\rho)^k}{c!n^{c-1}} \sum_{c'}^{c} \frac{t(k^4(2\rho)^t)}{n}.$$ 

Since $20 \log(2\rho) \log n - \log n \rightarrow \infty$, the largest term in the sum is when $c' = c$, and so our bound becomes

$$\frac{ct^{c+1}(2\rho)2^k k^{4c}}{c!n^{2c-1}}.$$
Dividing this expression by $E^2[T]$ we obtain

$$\frac{c4^e c^t t^{e+1} k^{4c}}{n} \leq \frac{k^{6c}}{n}.$$ 

Taking logarithms,

$$6c \log k - \log n \sim \frac{6 \log n}{13 \log n \log n} \log \log^2 n - \log n \sim \frac{12}{13} \log n - \log n \to -\infty.$$ 

Thus $E_2 = o(E^2[T])$, and we conclude $E[T^2] \sim E^2[T]$. Thus, $\text{Var}[T] = o(E^2[T])$, and so the number of these configurations approaches infinity whp. Therefore with high probability in a maximal matching there are at least $\frac{\log n}{13 \log \log n}$ row vertices not matched, for a total of twice this many row and column vertices not matched. □

We conclude our study of $B_2$ by giving a formula for the probability that $B_2$ is connected.

**Theorem 4.3.3.** The probability that $B_2$ is connected is given by

$$\prod_{j \geq 1} (1 - \lambda_j^2)^{1/2} \exp(\lambda_j^2/2),$$

where $\lambda_j$ are the eigenvalues of an integral operator with the kernel

$$K(x, y) = k(x) \land k(y), \quad k(x) = \exp(-x - e^{-\pi}(2 + x)).$$

Furthermore, this probability is bounded below by 0.996636.

**Proof.** Let $s_m(n)$ be the expected number of cycles of length $2m$ in $B_2$. Then

$$s_m(n) = \binom{n}{m} \binom{n}{m} \frac{(m-1)! m!}{2} \text{P}(S),$$
where $P(S)$ is the probability that the first $m$ vertices of $I$ and $J$ form the cycle
$(1, 1', 2, 2', \ldots, m, m')$, where $i \in I$ and $i' \in J$. Conditioning on the values in the
corresponding entries of the incidence matrix with weights $u_i = C_{i,i'}$, $u_i' = C_{i,(i-1)}$,
(taking $0' = m'$) and letting

$$X_i = \min \{ C_{i,j} : m + 1 \leq j \leq n \}$$

$$Y_j = \min \{ C_{i,j} : m + 1 \leq i \leq n \},$$

where $\min_2$ indicates second minimum, we have

$$P(S) = \mathbb{E} \left[ \prod_{i=1}^{m} e^{-\frac{1}{n} \sum_{j=m+1}^{n} \max(u_i, u_i')} \prod_{j=1}^{m} e^{-\frac{1}{n} \sum_{i=m+1}^{n} \max(u_i, u_i'-1, X_j)} \right] P(W),$$

(4.3.10)

where $W$ is the event that there are no other edges among the $2m$ vertices under
consideration.

From Theorem 4.3.1, we know that the smallest two entries in each of the first
$m$ rows are each one of the two smallest in one of the first $m$ columns. If there is an
entry besides those we are conditioning on that is smaller than all the others in the
last $n - m$ rows (or columns), then there is a set of $m$ rows (columns) with $2m + 1$
entries smaller than the entries, in their particular row (column), in the remaining
$n - m$ columns (rows). The expected number of such sets of rows and columns is
bounded by

$$2 \binom{n}{m} \binom{n}{m} m^{2m+1} \left( \frac{m}{n} \right)^{2m+1} \to 0,$$

since $m$ is fixed. Thus $W$ occurs whp.
Now, from (4.3.8) and (4.3.9), the probability that any of the exponents differs by more than \( n^{-1/4} \) of its expectation is exponentially small. From (4.3.6),

\[
E[\max(a, Z)] \approx a + (a + 2)e^{-a},
\]

with \( Z \) having the same distribution as our \( X \)'s and \( Y \)'s. Then we write

\[
P(S) \approx E \left[ \prod_{i=1}^{2m} e^{-u_i-(2+v_i)e^{-u_i}} \right],
\]

where \( v_{2i} = \max(u_i, u_i') \) and \( v_{2i-1} = \max(u_i, u_i'_{i-1}) \). Then we have

\[
s_m(n) \approx \binom{n}{m} \binom{n}{m} \frac{(m - 1)!m!}{2} \int \cdots \int_{\{u_1, \ldots, u_m, u_1', \ldots, u_m' \geq 0\}} 2^n e^{-u_i-(2+v_i)e^{-u_i}} \prod_{i=1}^{m} e^{-u_i/n} du_i \prod_{i=1}^{m} e^{-u_i'/n} du_i',
\]

so that

\[
s_m = \lim_{n \to \infty} s_m(n) = \frac{1}{2m} \int \cdots \int_{\{u_1, \ldots, u_m, u_1', \ldots, u_m' \geq 0\}} 2^n e^{-u_i-(2+v_i)e^{-u_i}} du_1 \cdots du_m du_1' \cdots du_m'.
\]

Now we introduce the symmetric integral operator \( K \) by

\[
(Kf)(x) = \int_0^\infty K(x, y)f(y)dy,
\]

\[
K(x, y) = \min(k(x), k(y)),
\]

\[
k(x) = \exp(-x - e^{-x}(2 + x)).
\]
Let $\lambda_1, \lambda_2, \ldots$ denote the eigenvalues of $K$. Then

$$\sum_{j \geq 1} \lambda_j = \int_0^\infty K(y, y) dy.$$ 

Now consider the operator $K^m$. It has kernel given by

$$K(m)(x, y) = \int \cdots \int_{\{y_1, \ldots, y_{m-1} \geq 0\}} K(x, y_1)K(y_1, y_2)\cdots K(y_{m-1}, y)dy_1\cdots dy_{m-1}.$$ 

Then, since $\lambda_1^{2m}, \lambda_2^{2m}, \ldots$ are the eigenvalues of $K^{2m}$, we have

$$\sum_{j \geq 1} \lambda_j^{2m} = \int_0^\infty K^{(2m)}(y_{2m}, y_{2m})dy_{2m}$$

$$= \int \cdots \int_{\{y_1, \ldots, y_{2m} \geq 0\}} K(y_{2m}, y_1)K(y_1, y_2)\cdots K(y_{2m-1}, y_{2m})dy_1\cdots dy_{2m}.$$ 

Notice that after dividing by $2m$, this last expression is equivalent to (4.3.11).

Therefore, if $s$ is the limiting expected number of cycles in $B_2$, then

$$s = \sum_{m \geq 2} s_m = \sum_{m \geq 2} \sum_{j \geq 1} \frac{1}{2m} \lambda_j^{2m}$$

$$= -\frac{1}{2} \sum_{j \geq 1} (\log(1 - \lambda_j^2) + \lambda_j^2/2).$$

Now, we can show that the number of cycles, $S$, in $B_2$ has a Poisson distribution. This follows because $E_r$, the $r$th factorial moment of $S$, converges to $E^r = E^r[S]$ (see [4]). Indeed, the only differences between these two expressions arise from:

(1) Binomial coefficients of the form $\binom{n}{m}$ in $E^r$ look like $\binom{n-m'}{m'}$, with $m'$ fixed, in $E_r$, (2) sums in exponents arising as in (4.3.10) have a negligible difference in the number of terms and (3) in $E^r$ there is a product of $P(W)$'s (corresponding to $P(W)$ in (4.3.10)) while in $E_r$ there is a corporate $W$, which also occurs whp.
Since the number of cycles has a Poisson distribution, the probability that there are no cycles is given by

\[ e^{-s} = \prod_{j \geq 1} (1 - \lambda_j^2)^{1/2} \exp(\lambda_j^2/2). \]

Notice that (4.3.11) may be bounded from above by replacing the \( u_i \) with one of the \( u_i \) or \( u'_i \). Doing this in one way results in

\[ s_m \leq \frac{1}{2m} \left( \int_0^\infty e^{-u-(2+u)} e^{-u} \, du \right)^{2m}. \]

By numerical integration and summing over \( m \) we obtain \( s \leq 0.0033696 \). Therefore, the probability that \( B_2 \) is connected is

\[ e^{-s} \geq e^{-0.0033696} \approx 0.996636. \]

4.4 Properties of \( B_k, k \geq 3 \)

Theorem 4.4.1. With high probability, \( B_3 \) has a perfect matching.

Proof. If there is not a perfect matching then by Hall's Lemma there must be a set of \( k \) row vertices which has at most \( k - 1 \) column vertex neighbors. Such an occurrence we will call a block to a perfect matching. The existence of a block of size \( k \) (i.e., having \( k \) row vertices) implies that there is also a set of \( n - k + 1 \) column vertices which has at most \( n - k \) row vertex neighbors, and so it suffices to consider only blocks of size \( k \leq n/2 \). Notice that in a block, in particular, each of the \( k \) rows must make its selections from the set of \( k - 1 \) column vertices. Let \( A_k \) be the
number of such sets of vertices. Then

\[
E(A_k) \leq \binom{n}{k} \left( \binom{n}{k-1} \left( \frac{k}{n} \right)^3 \right) \quad (4.4.1)
\]

\[
\leq \frac{n^{2k} k^{3k}}{(k!)^2 n^{3k}} \leq \left( \frac{ke^2}{n} \right)^k. \quad (4.4.2)
\]

Thus, if \( P_k \) denotes the probability that there is a block of size \( k \), then the probability that there is a block of size less than \( t = \lfloor en \rfloor \) is, letting \( \ell = \lfloor \log \log n \rfloor \),

\[
\sum_{k=4}^{t} P_k \leq \sum_{k=4}^{\ell} \left( \frac{ke^2}{n} \right)^k + \sum_{k=\ell+1}^{t} \left( \frac{ke^2}{n} \right)^k 
\]

\[
\leq \sum_{k=4}^{\ell} \left( \frac{e^2e^2}{n} \right)^k + \sum_{k=\ell+1}^{t} (ee^2)^k 
\]

\[
\leq \frac{e^{2e^2}}{n} + 2(ee^2)^\ell \to 0,
\]

if \( e < \frac{1}{2e^2} \).

Now we’ll show that for any \( k \) of order \( n \) there are no blocks to a perfect matching. For this we will need to use the fact that in any block there is also a set of \( n - k + 1 \) columns that do not chose any of the \( k \) rows.

Suppose the vertices of a block label the first \( k \) rows and \( k - 1 \) columns of the incidence matrix. Then the upper right corner of the matrix consists of a submatrix of size \( k \times (n - k + 1) \) in which no edges are found. To calculate the expected number of blocks then amounts to calculating the expected number of submatrices with this property. Consider a submatrix of size \( k \times (n - k + 1) \). Let \( Y_i \) be the smallest entry in row \( i \) of this submatrix and \( Z_j \) the smallest entry in column \( j \) of this submatrix. Each \( Y \) has an exponential distribution with parameter \((n - k + 1)/n\), and each
$Z$ has an exponential distribution with parameter $k/n$. Notice that the submatrix will contain no edges if each row and column of the incidence matrix corresponding to a row or column of the submatrix has 3 entries smaller than the corresponding $Y$ or $Z$. If we condition on the value of $Y_i$, then the probability that row $i$ contains at least 3 entries smaller than $Y_i$ is given, letting $y = Y_i$, by

$$
P_{n,k}(y) = 1 - (e^{-y/n})^{k-1} - k(1 - e^{-y/n})(e^{-y/n})^{k-2}$$
$$- \binom{k}{2}(1 - e^{-y/n})^2(e^{-y/n})^{k-3},$$

since the complimentary event has exactly zero, one or two of the $k - 1$ first entries of row $i$ smaller than $Y_i$.

Since we are only concerned now with $k$ of order $n$, we let $a = k/n$ and write

$$P_{n,k}(y) \sim f(a,y) = 1 - e^{-ay} - ay e^{-ay} - \frac{a^2y^2}{2} e^{-ay}.$$  

Similarly, the probability that there are three entries in column $j$ smaller than $Z_j$ is $f(1 - a, Z_j)$.

Now we can bound the probability that there is a block to a perfect matching of size $k$ by

$$P_k \leq \binom{n}{k} \binom{n}{k-1} E_k.$$
where

\[ E_k = E \left( \prod_{i=1}^{k} f(a, Y_i) \prod_{j=k}^{n} f(1-a, Z_j) \right) \]

\[ \leq E^{1/2} \left( \prod_{i=1}^{k} f^2(a, Y_i) \right) E^{1/2} \left( \prod_{j=k}^{n} f^2(1-a, Z_j) \right) \]

\[ = \prod_{i=1}^{k} E^{1/2}(f^2(a, Y_i)) \prod_{j=k}^{n} E^{1/2}(f^2(1-a, Z_j)) \]

\[ = E^{k/2}(f^2(a, Y)) E^{(n-k+1)/2}(f^2(1-a, Z)). \]

Asymptotically, \( Y \) has an exponential \((1-a)\) distribution, so we can compute

\[ E[f^2(a, Y)] = \int_{0}^{\infty} (1 - e^{-ay} - aye^{-ay} - \frac{a^2y^2}{2}e^{-ay})^2(1-a)e^{-(1-a)y} \, dy \]

\[ = (1-a) \int_{0}^{\infty} e^{-(1+a)y}(e^{ay} - 1 - ay - a^2y^2/2)^2 \, dy \]

\[ = (1-a) \int_{0}^{\infty} e^{-(1+a)y}(e^{2ay} - 2e^{ay} - 2aye^{ay} - a^2y^2e^{ay} + 1 \]

\[ + 2ay + 2a^2y^2 + a^3y^3 + a^4y^4/4) \, dy \]

\[ = (1-a) \int_{0}^{\infty} (e^{-(1-a)y} - 2e^{-y} - 2aye^{-y} - a^2y^2e^{-y} + e^{-(1+a)y} \]

\[ + 2aye^{-(1+a)y} + 2a^2y^2e^{-(1+a)y} + a^3y^3e^{-(1+a)y} + a^4y^4e^{-(1+a)y}/4) \, dy. \]

Now using

\[ \int_{0}^{\infty} y^i e^{-cy} \, dy = \frac{i!}{c^{i+1}} \]

we obtain,

\[ E[f^2(a, Y)] = (1-a) \left( \frac{1}{1-a} - 2 - 2a - 2a^2 + \frac{1}{1+a} \right. \]

\[ + \frac{2a}{(1+a)^2} + \frac{4a^2}{(1+a)^3} + \frac{6a^3}{(1+a)^4} + \frac{6a^4}{(1+a)^5} \]

\[ \left. = 2a^6(10 + 5a + a^2) \right\} \frac{1}{(1+a)^5}. \]
Letting \( g(a) = \mathbb{E}^{a/2}[f^2(a, Y)] \), we have

\[
P_k \leq \binom{n}{k} \binom{n}{k-1} (g(a)g(1-a))^n \sim \left( \frac{1}{a^{2a}} \frac{1}{(1-a)^{2(1-a)}} g(a)g(1-a) \right)^n = h^n(a).
\]

A plot of \( h(a) \) shows that the function has a local maximum of \( \approx .916 \) at 0.5 and then approaches 1 as \( a \to \pm 1 \). In any case, it is bounded below 1 for \( a \) in \([\delta, 0.5]\), which implies that the expected number of blocks to a perfect matching is exponentially small for any \( k > \delta n \) (with \( k < n/2 \)). Taking \( \delta < \varepsilon \) chosen earlier, we conclude that there are no blocks to a perfect matching, with high probability.

Now we can use the properties we've discovered of \( B_2 \) and \( B_3 \) to prove the following.

**Theorem 4.4.2.** With high probability, \( B_3 \) is connected.

**Proof.** Because \( B_3 \) has a perfect matching whp, each component has the same number of vertices from each part of the bipartite graph whp. Now notice that the bound in (4.4.2) holds even when \( k \) replaces \( k - 1 \) in the second binomial coefficient of (4.4.1). This implies that there are no components of size less than \( \varepsilon n \) for some fixed \( \varepsilon \). Finally, since \( B_2 \) contains a single component with size of order \( n \) whp, the same must be true of \( B_3 \). Thus \( B_3 \) consists of one component whp.
4.5 Open problems

A major question that remains unanswered is how many vertices can be matched in $B_2$ whp. Computer experiments suggest that the fraction of vertices matched approaches 1 whp. We are continuing to attack the problem from two angles. One of these uses the idea of a minimal configuration, similar to what we did in the proof of Theorem 4.3.1. The goal is to describe a structure that must be present if the graph does not have a matching of at least $n - f(n)$ of the vertices, and show that the expected number of these goes to zero. We would then conclude that there is a matching with at least $n - f(n)$ vertices whp. The other idea is to find an algorithm that successfully matches $n - f(n)$ of the vertices whp. We have tried several greedy algorithms, but found that they all match only about 80% of the vertices. Recall that we were able to show that $B_1$ already has a matching with about 80% of the vertices whp.

This last fact suggests that we could do better by starting with a matching in $B_1$ and extend it to a matching in $B_2$. Successfully doing so will not only help answer our first question, but should also help answer some questions concerning $B_3$.

We know that $B_3$ contains a perfect matching whp. What is the expected weight of one of these matchings? It seems likely that there are many perfect matchings in $B_3$, especially if we are right in thinking that $B_2$ contains a matching with $n - o(n)$ edges whp. Allowing the computer to find a perfect matching in $B_3$ "from scratch" with $n = 1000$ and with the $C_{i,j}$ uniform $(0, 1)$ random variables results in matchings with weights approximately 2.5. On the other hand, if we start with
matchings in $B_1$, extend to matchings in $B_2$ and then finally to matchings in $B_3$ we obtain matchings with weights around 1.9, which is less than the current upper bound of 2 (proved by Karp [18]). It seems clear, then, that an understanding of the relationships between the $B_k$ is central to answering important questions concerning near optimal matchings in these graphs.
LIST OF REFERENCES


