

# **Applications of the Probabilistic Method to Random Graphs**

by

**Vishal Sanwalani**

B.A., Finance, Economics, and Biology, Washington University, 1997

DISSERTATION

Submitted in Partial Fulfillment of the  
Requirements for the Degree of

Doctor of Philosophy  
Computer Science

The University of New Mexico

Albuquerque, New Mexico

July, 2005

©2005, Vishal Sanwalani

# Acknowledgments

I would like to thank my advisor, Professor Cristopher Moore, for giving me the freedom to learn. I would like to thank Professor Jared Saia for exposing me to new areas in Computer Science. I would like to thank my committee members Professor Shuang Luan and Professor Vladimir Koltchinskii for their helpful comments and suggestions. I would like to thank Professor Josep Diaz for all the valuable advice and help he has given me. Finally, I would like to thank all my coauthors, much of the work in this thesis was done jointly with their help.

# **Applications of the Probabilistic Method to Random Graphs**

by

**Vishal Sanwalani**

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the  
Requirements for the Degree of

Doctor of Philosophy  
Computer Science

The University of New Mexico

Albuquerque, New Mexico

July, 2005

# Applications of the Probabilistic Method to Random Graphs

by

**Vishal Sanwalani**

B.A., Finance, Economics, and Biology, Washington University, 1997

Ph.D., Computer Science, University of New Mexico, 2005

## Abstract

We discuss three problems. Although the problems themselves are distinct, two basic themes underly each of them. First, each problem is either directly, or can be viewed as a random graph problem. Second, all the problems can be solved by using elementary techniques from the probabilistic method. We first discuss the Leader Election problem. In the leader election problem, there are  $n$  processors,  $\beta n$  of which are *bad* (or corrupt), and  $(1 - \beta)n$  of which are *good*, for some fixed  $\beta$ . For  $\beta < 1/3$ , we present an algorithm which elects a leader from the set of all processors such that, with constant probability, this leader is good, and a  $1 - o(1)$  fraction of the good processors know the election result. Further the algorithm only requires each good processor to send and process a number of bits which is polylogarithmic in  $n$ .

Next we discuss the chromatic number of a random scaled sector graph. In the random scaled sector graph model, vertices are placed uniformly at random into the  $[0, 1]^2$  unit square. Each vertex  $i$  is assigned a uniformly at random sector  $S_i$ , of central angle  $\alpha_i$ , in a circle of radius  $r_i$  (with vertex  $i$  as the origin). An arc is present from vertex  $i$  to any vertex  $j$ , if  $j$  falls in  $S_i$ . We study the value of the chromatic number  $\chi(G_n)$ , for random scaled sector graphs with  $n$  vertices,

where each vertex spans a sector of  $\alpha$  degrees with radius  $r_n = \sqrt{\frac{\ln n}{n}}$ . We prove that for values  $\alpha < \pi$ , as  $n \rightarrow \infty$  w.h.p.,  $\chi(G_n)$  is  $\Theta(\frac{\ln n}{\ln \ln n})$ . For  $\alpha > \pi$  w.h.p. (with high probability),  $\chi(G_n)$  is  $\Theta(\ln n)$ .

Finally, we discuss the probability a sparse random graph or hypergraph is connected. While it is exponentially unlikely that a sparse random graph or hypergraph is connected, with probability  $1 - o(1)$  such a graph has a “giant component” that, given its numbers of edges and vertices, is a *uniformly* distributed connected graph. This simple observation allows us to estimate the number of connected graphs, and more generally the number of connected  $d$ -uniform hypergraphs, on  $n$  vertices with  $((d - 1)^{-1} + \Omega(1))n \leq m = o(n \ln n)$  edges.

# Contents

<b>List of Figures</b>	<b>xii</b>
<b>1 Scalable Leader Election</b>	<b>1</b>
1.1 Acknowledgements . . . . .	1
1.2 Introduction . . . . .	1
1.2.1 Problem Statement . . . . .	2
1.2.2 Our Results . . . . .	3
1.2.3 Related Work . . . . .	4
1.2.4 Roadmap . . . . .	5
1.3 Preliminaries . . . . .	5
1.4 The Layered Network . . . . .	9
1.5 Communication and validation . . . . .	10
1.5.1 Monitoring sets . . . . .	10
1.5.2 Validation between monitoring sets . . . . .	11

## Contents

1.5.3	Downward communication tree . . . . .	13
1.5.4	The Communications Protocol . . . . .	13
1.6	The Leader Election Algorithm . . . . .	14
1.7	Proof of Theorem 1 . . . . .	15
<b>2</b>	<b>The chromatic number of random scaled sector graphs</b>	<b>21</b>
2.1	Acknowledgements . . . . .	21
2.2	Introduction . . . . .	21
2.3	Results . . . . .	23
2.4	Basic constructions and lemmas . . . . .	24
2.5	Proof of Theorem 1 . . . . .	27
2.5.1	$\alpha < \pi - \epsilon$ . . . . .	27
2.5.2	$\alpha > \pi + \epsilon$ . . . . .	27
2.5.3	$\alpha = 2\pi$ . . . . .	32
2.6	Proof of Theorem 2 . . . . .	32
2.6.1	$\alpha > \pi + \epsilon$ . . . . .	32
2.6.2	$\epsilon < \alpha < \pi - \epsilon$ . . . . .	33
2.7	Proof of Theorem 3 . . . . .	36
2.7.1	$\alpha > \pi + \epsilon$ . . . . .	36
2.7.2	$\alpha < \pi - \epsilon$ . . . . .	37
2.8	Conclusions and open problems . . . . .	37

<b>3</b>	<b>Counting Connected Graphs and Hypergraphs via the Probabilistic Method</b>	<b>42</b>
3.1	Acknowledgements . . . . .	42
3.2	Introduction and Results . . . . .	42
3.2.1	Results . . . . .	43
3.2.2	Techniques and Overview . . . . .	47
3.2.3	Related Work . . . . .	49
3.3	Preliminaries . . . . .	51
3.4	The Number of Connected Hypergraphs . . . . .	53
3.4.1	Outline . . . . .	53
3.4.2	Proof of Lemma 2 . . . . .	58
3.4.3	Proof of Lemma 3 . . . . .	58
3.4.4	Proof of Lemma 4 . . . . .	60
3.5	The Probability of Getting a Giant Component of a Given Order and Size . . . . .	61
3.5.1	Outline . . . . .	61
3.5.2	Proof of Lemma 7 . . . . .	66
3.5.3	Proof of Lemma 8 . . . . .	67
3.5.4	Proof of Lemma 9 . . . . .	68
3.6	The Expected Number of Edges Given that $H_d(n, p)$ is Connected . . . . .	70
3.7	Branching Processes and The Giant Component of $H_d(n, p)$ . . . . .	71
3.7.1	Preliminaries on Branching Processes . . . . .	72

*Contents*

3.7.2	Exploring the Components of $H_d(n, p)$ . . . . .	76
3.7.3	Large Deviations of $\mathcal{N}(H_d(n, p))$ and the Number of Isolated Vertices . .	83
3.7.4	The Variance of $\mathcal{N}(H_d(n, p))$ . . . . .	85
3.7.5	The Variance of the Number of Edges Outside the Giant . . . . .	88
3.7.6	Proof of Lemma 14 . . . . .	93

<b>References</b>		<b>95</b>
-------------------	--	-----------

# List of Figures

2.1	The sector of a sensor $i$ and the communication between motes . . . . .	23
2.2	Angle partition for $\alpha > \pi + \epsilon$ (a) classes $\mathcal{B}$ (b) directions associated to a class $B_j$	25
2.3	The basic dissections of $[0, 1]^2$ (a) $\mathcal{S}$ (b) horizontal subdivision (c) vertical subdivision . . . . .	26
2.4	Proof of lower bound . . . . .	39
2.5	Partition of $S$ by strips . . . . .	40
2.6	Sector $S_i$ and complementary sector $S_i^*$ . . . . .	40
2.7	Figure for the proof of 5.2 . . . . .	41
2.8	Partition of $S$ in the prove 5.2 . . . . .	41

# Chapter 1

## Scalable Leader Election

### 1.1 Acknowledgements

I would like to thank Valerie King, Jared Saia, and Erik Vee. The work in this chapter was done jointly with them.

### 1.2 Introduction

Leader election is a fundamental problem in distributed computing. We consider this problem in the setting where there are  $n$  processors,  $\beta n$  of which are *bad* (or corrupt), and  $(1 - \beta)n$  of which are *good*, for some fixed  $\beta$ . We assume that an omniscient and computationally unbounded adversary picks which processors will be bad before the algorithm begins and this adversary controls the actions of all bad processors so as to maximize the chance of getting a bad processor elected. Our goal is to design an algorithm which ensures a good processor will be elected with constant probability, no matter which set of  $\beta n$  processors are bad. This problem was first formally described and addressed by Ben-Or and Linial [4, 5] about twenty years ago. Since then,

many papers have been published giving leader election algorithms that successively improve on the number of rounds required and on the fraction  $\beta$  of bad processors that can be tolerated [18, 1, 2, 7, 6, 9, 15, 17, 12] (see also surveys by Ben-Or, Linial and Saks [3] and Linial [14]).

In this chapter, we describe a leader election algorithm which is *scalable*, in the sense it requires each good processor to send and process a number of bits which is polylogarithmic in  $n$ .

### 1.2.1 Problem Statement

The standard model for communication in the leader election problem is the *full information model*<sup>1</sup> [4, 5]. In this model, all communication occurs by broadcast and is known publicly to all processors. Every processor has a unique name known to everyone and the name of the sender of any message is explicitly known by all processors. Communication occurs in *rounds*. In each round, every processor may communicate with all other processors. The bad processors are assumed to have received the messages of all the good processors before they broadcast their own messages. The processors are synchronized between rounds so that all messages in round  $i$  are assumed to be received before any messages in round  $i + 1$  are sent out. Since the adversary is computationally unbounded, this disallows the use of cryptographic assumptions.

The full information model rules out from the very start any possibility of designing an algorithm where each processor sends a sublinear number of bits. In particular, since all communication is by broadcast, every time a node communicates in this model, it sends out  $\Omega(n)$  bits. To get around this problem, we use the *point-to-point full information model*. In this model, all communication occurs between a single sender and a single receiver. The bad processors see all messages but the good processors only see messages that are sent directly to them. Everything else is the same as in the full information model. In particular, each processor has a unique name which is known explicitly to anyone to whom it sends messages. Further, communication occurs

---

<sup>1</sup>This is sometimes also referred to as the *perfect information model*.

in rounds and the bad processors see all messages before they need to send their own. This new model is strictly harder than the standard full information model in the following sense. In the standard model, in a single round, a bad processor is forced to send the same message to all processors (since communication is by broadcast). However, in the new model, a bad processor can send different messages to different processors.

Our goal is to minimize the number of bits *sent* and *processed* by every good processor. We assume that a processor can choose to ignore (not process), without cost, messages received from any other processor during any round of the algorithm.

## 1.2.2 Our Results

We conjecture, if a constant fraction of the processors are bad, then any algorithm which insures with positive probability a good leader is elected and all good processors know the leader, will require each processor to send and process  $\Omega(n)$  bits. Thus, we relax the requirement that *all* good processors know the leader at the end of the protocol to the requirement that a  $1 - o(1)$  fraction of good processors know the leader. Our main result is stated as Theorem 1 below. To the best of our knowledge, this is the first result for the leader election problem where each processor is required to send and process a sub-linear number of bits.

**Theorem 1** *Assume there are  $n$  processors and strictly less than a  $1/3$  fraction of these processors are bad. Then there exists an algorithm that elects, with constant probability, a leader from the set of good processors and has the following properties.*

- *Exactly one good processor considers itself the leader.*
- *A  $1 - o(1)$  fraction of the good processors know this leader.*
- *Every good processor sends and processes only a polylogarithmic (in  $n$ ) number of bits.*
- *The number of rounds required is polylogarithmic in  $n$ .*

### 1.2.3 Related Work

The first results for leader election in the full information model are due to Ben-Or and Linial [4, 5]. They give a one round protocol which is robust to up to a  $1/\ln n$  fraction of bad processors. Kahn, Kalai and Linial [13] show that if each good processor is restricted to providing one random bit, then  $1/\ln n$  is the largest fraction of bad processors that can be tolerated for any one round protocol. Saks [18] and Ajtai and Linial [1] designed “baton-passing” protocols which are robust to a  $1/\ln n$  fraction of bad processors. Saks [18] also showed that no protocol can be robust against  $\lceil n/2 \rceil$  bad processors. Alon and Naor [2] designed a modified baton-passing protocol which they showed to be robust to  $\beta n$  bad processors for any  $\beta < 1/3$ . Further analysis of Alon and Naor’s protocol by Boppana and Narayanan [7, 6] showed that the protocol was robust to  $\beta n$  bad processors for all  $\beta < 1/2 - \epsilon$  and positive  $\epsilon$ .

The protocol due to Alon and Naor has optimal resilience but requires a linear number of rounds. Several subsequent papers focused on reducing the number of rounds [9, 15, 19, 17]. Russell and Zuckerman [17] designed a protocol which is resilient against  $(1 - \epsilon)n/2$  bad processors and takes only  $\ln^* n$  rounds. Russell, Saks and Zuckerman [16] further showed that  $\Omega(\ln^* n)$  rounds are necessary if in every round the good processors each send one unbiased random bit. Finally, Feige [12] designed a much simpler protocol which requires  $O(\ln^* n)$  rounds and in the situation where there are  $(1 + \delta)n/2$  good processors, elects a good processor with probability  $\Omega(\delta^{1.65})$ . All previous protocols had success probability which was exponentially small in  $\delta$ .<sup>2</sup> All of these algorithms require each processor to send and process  $\Omega(n)$  bits, yet it is difficult to directly compare our result with previous algorithms. First, the previous algorithms assumed messages were sent via broadcast, hence the focus was on reducing the number of rounds required. Second, we have relaxed the assumption every good processor knows the leader (if messages are broadcast this assumption is trivially satisfied).

---

<sup>2</sup>The success probability was still constant provided that  $\delta$  was constant.

## 1.2.4 Roadmap

In section 1.3 we detail the lemmas, definitions, and previous results which will be used in the remainder of the chapter. In sections 1.4 and 1.5 we describe the constructions and protocols which will be used in the algorithm. In section 1.6 we described the Leader Election Algorithm. Our proof of Theorem 1 is presented in section 1.7.

## 1.3 Preliminaries

We will use the phrase *with high probability* (or simply *w.h.p.*) to mean that an event happens with probability at least  $1 - o(n^{-c})$  for some  $c > 2$ . For readability, we treat  $\ln n$  as an integer.

We adapt an algorithm by Feige [12] to the point-to-point full information model to get what we will call the *heavy weight leadership election protocol*. This algorithm gives a heavy weight method for electing a good leader with constant probability from among a set of  $n$  processors, among which a fraction greater than  $2/3$  are good. Feige's result shows that this can be done in  $\ln^* n$  expected rounds, with each processor broadcasting at most once per round. To adapt Feige's result to the point-to-point full information, we replace each broadcast operation with a call to a Byzantine Agreement algorithm such as [10] (the algorithm presented in [10] requires  $O(n)$  rounds and  $O(n^3)$  bits when there are  $n$  processors, and less than  $n/3$  are bad). This results in an algorithm which requires  $O(n^4 \ln^* n)$  bits.

**Lemma 1** [12] *In the point-to-point full information model, there is a heavy weight leadership election protocol with the following properties. On a set of  $n$  processors, with  $(2/3 + \epsilon)n$  good processors, it returns a good leader with probability at least  $\Omega((1/3 + 2\epsilon)^{1.65})$  and requires  $O(n^4 \ln^* n)$  bits and  $O(n^2 \ln^* n)$  rounds.*

Next, we describe a method for electing a subcommittee of processors, with desired properties, from a committee of processors using what we call the *subcommittee election protocol*.

Chapter 1. Scalable Leader Election

This protocol is also a simple adaptation of an algorithm in [12]:

1. Given a committee of processors  $p_i$  in  $S$ , where for some constant  $C_1$ ,  $|S| = C_1 \ln^8 n$ . For  $i = 1, \dots, |S|$ , do the following:
  - (a) Processor  $p_i$  randomly selects one of  $\ln^5 n$  “bins” and writes the other processors in its committee with the bin it has selected
  - (b) The other processors in  $S$  do Byzantine Agreement to come to consensus on which bin  $p_i$  has selected
2. Let  $B$  be the bin with the least number of processors in it and let  $S_B$  be the set of processors in that bin. Pad the set of leaders selected in the set  $S_B$  with enough additional arbitrary processors added to ensure  $|S_B| = C_1 \ln^3 n$ . The processors in  $S_B$  represent the elected subcommittee.

**Lemma 2** *Let  $S$  be a committee of processors, where the fraction,  $f_S$ , of good processors is  $> 2/3$ . Then for any constant  $c \geq 2$ , there exists a constant  $C_1$ , such that with probability at least  $1 - 1/n^c$ , the subcommittee election protocol elects a subset  $Z$  of  $S$  with the following property.  $|Z| = C_1 \ln^3 n$  and the fraction of good processors in  $Z$  is greater than  $(1 - 1/\ln n)f_S$ . Further, this algorithm uses a polylogarithmic number of bits and polylogarithmic number of rounds.*

**Proof** Let  $X$  be the smallest number of good processors in any bin. Define  $P_1$  to equal  $Pr[X < (1 - 1/\ln n)f_S C_1 \ln^3 n]$ . The number of good processors in any one bin is a binomially distributed random variable, with mean  $\mu = f_S C_1 \ln^3 n$ . Thus using Boole’s inequality and the Chernoff bound on the binomial distribution,  $P_1 < n \cdot \exp(-f_S C_1 \ln n/3)$ .

Where in the above equation we have (generously) bounded the number of bins by  $n$ . Since  $f_S > 1/2$ , setting  $C_1 = 6(c + 1)$ , we have  $P_1 < 1/n^c$ . Since by construction  $|S_B| = C_1 \ln^3 n$ , we have established the first part of Lemma 2.

## Chapter 1. Scalable Leader Election

We now establish a polylogarithmic bound on the number of bits and rounds used in sub-committee election protocol. We first note the message and round cost of the algorithm are both polynomial in the number of processors participating in the algorithm. We finally note that by assumption,  $|S|$ , the number of processors participating in the algorithm, is  $\Theta(\ln^8 n)$ .  $\square$

Next, we present a result similar to one used in [9]. Let  $X$  be a set of processors. For a family  $F$  of subsets of  $X$ , a parameter  $\delta = 1/\ln n$ , and a subset  $X' \subseteq X$ , let  $F(X', \delta)$  be the family of all  $F' \in F$  for which

$$\frac{|F' \cap X'|}{|F'|} < \frac{|X'|}{|X|} + \delta.$$

Let  $\Gamma(r)$  denote the neighbors of node  $r$  in a graph.

**Lemma 3** *For some constant  $c$ , we can create a family of bipartite graphs  $G(L_i, R_i)$ ,  $i = 0, 1, 2, \dots, \ln n / \ln \ln n - c$  where for all  $i$ ,  $|L_i| = n / \ln^i n$ ;  $|R_i| = n / (C_1 \ln^{i+4} n)$ , the degree of all nodes in  $R_i$  is  $C_1 \ln^8 n$ ; such that, if we set  $X = L_i$  and  $F = \{\Gamma(r) \mid r \in R_i\}$ , then:*

- *For any subset  $X' \subseteq X$ ,  $F(X', \delta) < |X| / \ln^6 n$ .*
- *No node  $l \in L_i$  have multiple edges with any node  $r \in R_i$ .*
- *Each node  $l \in L_i$  has degree at most  $2 \ln^4 n$ .*

**Proof** Consider a bipartite graph  $G(L, R)$  where there is a node in  $L$  for each element of  $X$  and  $|X| / (C_1 \ln^4 n)$  nodes in  $R$ , one for each element of  $F$ . An edge between a node  $l \in L$  and a node  $r \in R$  corresponds to the element represented by  $l$  appearing in the set represented by  $r$ ; each node in  $R$  has degree  $C_1 \ln^8 n$ . We will show, if for each  $r \in R$  its neighbors in  $L$  are selected independently and uniformly at random without replacement, then w.h.p.,  $G(L, R)$  has the properties specified in the lemma. We note, since the neighbors for each  $r \in R$  are selected without replacement, the second property of the lemma is trivially satisfied.

## Chapter 1. Scalable Leader Election

Now we prove  $G(L, R)$  will have the first property described in the lemma with probability greater than  $1 - 1/n$ . Let  $G(L, R)$  be a random bipartite graph as specified above, let  $L'$  be some fixed subset of  $L$  and let  $R'$  be some fixed subset of  $R$  where  $|R'| = |X|/\ln^6 n$ . Let  $\xi(L', R')$  be the event that every node in  $R'$  has more than a  $|L'|/|L| + \delta$  fraction of its edges incident to nodes in  $L'$ . We now show that the probability of this event is very small.

Let  $N(L', R')$  be the number of edges between  $L'$  and  $R'$ . Note that the number of edges incident (from  $L$ ) to nodes in  $R'$  is  $(C_1 \ln^8 n)|R'|$ . Thus for  $\xi(L', R')$  to occur, it must be the case that  $N(L', R') > (|L'|/|L| + \delta)((C_1 \ln^8 n)|R'|)$ . By linearity of expectation,  $E(N(L', R')) = (|L'|/|L|)((C_1 \ln^8 n)|R'|)$ . Since the edges incident to  $R'$  are chosen via  $(C_1 \ln^8 n)|R'|$  trials, and if any one trial is altered, the expected effect on  $N(L', R')$  is bounded above by 2; by Azuma's inequality,  $Pr(\xi(L', R')) \leq \exp(-\delta^2(C_1 \ln^8 n)|R'|/8)$ .

Now let  $P_2$  be the probability that there exists any  $L' \subseteq L$  and any  $R' \subseteq R$  such that  $|R'| = |X|/\ln^6 n = |L|/\ln^6 n$  and the event  $\xi(L', R')$  occurs. By Boole's inequality:

$$\begin{aligned} P_2 = Pr\left(\bigcup_{L' \subseteq L, R' \subseteq R} \xi(L', R')\right) &\leq \sum_{L' \subseteq L, R' \subseteq R} Pr(\xi(L', R')); \\ &\leq 2^{|L|} 2^{|R|} \exp(-\delta^2(C_1 \ln^8 n)|R'|/8); \\ &\leq \exp(|L| + |R| - C_1|X|/8); \\ &\leq 1/n^{C_1/8-2}. \end{aligned}$$

Where the second to last line follows since  $\delta^2|R'| \geq |X|/\ln^8 n$ ; and the last line follows since  $|R| < |L| = |X|$ , and  $|X| \geq \ln^8 n$ . Setting  $C_1 = 24$ , shows, for sufficiently large  $n$ ,  $P_2 < 1/n$

Finally, we prove  $G(L, R)$  will have the third property described in the lemma with probability greater than  $1 - 1/n$ . Let  $P_3$  be the probability the degree of any node  $l \in L$  is greater than  $2 \ln^4 n$ . By linearity of expectation, the average degree of each node  $l \in L$  is  $\ln^4 n$ . Since

the neighbors for every node  $r \in R$  are chosen independently of each other, by applying the Chernoff bound and Boole's inequality,  $P_3 < n \cdot e^{-\ln^4 n/3} < 1/n$ , for  $n$  sufficiently large.

Noting  $P_2 + P_3 < 2/n$ , and the total number of bipartite graphs is  $O(\ln n)$  completes the proof.  $\square$

## 1.4 The Layered Network

We construct a layered network. The index  $i^*$  of the top layer is the minimum integer  $i^*$  such that  $n/\ln^{i^*} n < \ln^{10} n$ . The nodes in the layered network will correspond to committees. Where a committee is a collection of processors. In order to avoid confusion we will refer to the nodes in the network as *committee nodes*. All committee nodes, excluding the committee node on the top layer, consist of  $C_1 \ln^8 n$  processors. Initially, in order to assign the  $n$  processors to committee nodes on layer 0 we use the bipartite graph  $G(L_0, R_0)$  described in Lemma 3. Each node  $r \in R_0$  represents a committee node on layer 0, and each of the  $n$  processors is represented by a node  $l \in L_0$ . The layer 0 committee node represented by a node  $r \in R_0$  consists of the set of processors represented by its neighbors in  $G(L_0, R_0)$ .

Processors are assigned to the committee nodes on layer  $i + 1$  as follows. In parallel, the processors in each layer  $i$  committee node,  $l$ , hold an election using the subcommittee election protocol to elect a subcommittee of processors. When a processor is elected to a subcommittee on layer  $i$ , we will refer to the processor as being *elected* on layer  $i$ . In the bipartite graph  $G(L_{i+1}, R_{i+1})$  described in Lemma 3; each node  $r \in R_{i+1}$  represents a committee node on layer  $i + 1$ , and each of the elected processors on layer  $i$  is represented by a node  $l \in L_{i+1}$ . The layer  $i + 1$  committee node represented by a node  $r \in R_{i+1}$  consists of the set of processors represented by its neighbors in  $G(L_{i+1}, R_{i+1})$ . If a processor elected in a committee node  $A$  on layer  $i$  is assigned to a committee node  $B$  on layer  $i + 1$  we will represent this by an edge between committee nodes  $A$  and  $B$  in the layered network. Further we will say that  $A$  is a child

of  $B$ . When the number of elected processors elected on layer  $i$  is less than  $\ln^{10} n$  all the elected processors are assigned to the single committee node on layer  $i^*$ , and these processors hold a leader election using the heavy weight leader election protocol referred to in Lemma 1.

**Observation 1** *At each layer  $i < i^*$ , there are  $n/(C_1 \ln^{i+4} n)$  committee nodes, and  $n/\ln^i n$  processors. The number of layers,  $i^* + 1$ , in the network is  $O(\ln n / \ln \ln n)$ .*

## 1.5 Communication and validation

As processors move up the layered network, the results of an election are not necessarily known to the other nonparticipating processors. We provide such information on a need-to-know basis, by establishing *monitoring sets*, one for each committee node election, and a communication tree to communicate the election results to the monitoring sets.

One interesting aspect of this problem is that straightforward polling can be defeated by flooding. That is, suppose a majority of processors in a monitoring set are good and correctly know an election result. The strategy of requesting the result from a random subset of that set may be thwarted because the bad processors may swamp the good processors with similar requests. The good processors, having a limit on the number of bits they may send, need to know which requests to ignore.

### 1.5.1 Monitoring sets

We create a monitoring set for each committee node election. The assignment of processors to monitoring sets is predetermined at the start of the algorithm. The monitoring sets for the elections on layer 0 are the processors in the committee nodes involved in those elections. Let  $z$  be the number of committee nodes (i.e. elections) on layer  $i$ . For each layer  $i > 0$ , the monitoring sets for the committee node elections on layer  $i$  are determined by an arbitrary partition of the

$n/(C_1 \ln^4 n)$  layer 0 committee nodes into  $z$  classes of equal size. So, for example, order the committee nodes (1 through  $z$ ) on layer  $i$ . Then the processors in committee nodes 1 through  $n/(zC_1 \ln^4 n)$  on layer 0, monitor the election of committee node 1 on layer  $i$ , etc. Thus processors in a monitoring set know the identities of the processors in the subcommittee elected on layer  $i$ , and hence those processors sent to the committee nodes on layer  $i + 1$ .

### 1.5.2 Validation between monitoring sets

In the course of the algorithm the processors elected in committee node  $A$ , may need to know the identities of the processors elected in committee node  $B$ , and we represent this by the pair  $(A, B)$ . For example if some of the elected processors in both  $A$  and  $B$  are assigned to the same the committee node (one layer above them), then we have  $(A, B)$  and  $(B, A)$ , since the processors elected to both subcommittees need to know the identities of each other. In our algorithm we will insure all such pairs  $(A, B)$  are predetermined by the layered network and committee nodes  $A$  and  $B$  are separated by at most one layer in the network. Let  $m(A)$  and  $m(B)$  denote the monitoring sets for  $A$  and  $B$  respectively. As noted previously, the processors in  $m(B)$  need to know the processors elected in  $A$ , otherwise the adversary can flood  $m(B)$ . Thus before the set  $m(B)$  is polled its processors must validate the elected subcommittee of  $A$  by polling  $m(A)$ . Thus each polling step will take place over two rounds, the first round being a validation phase.

Since the monitoring sets are growing as processors move up the network, each processor in  $m(B)$  can only poll a subset of  $m(A)$ . Recalling each monitoring set consists of layer 0 committee nodes, we define  $|m(A)|$  and  $|m(B)|$  to be the number of committee nodes in  $m(A)$  and  $m(B)$  respectively. Let  $x = |m(A)|$ , and  $y = |m(B)|$ ; without loss of generality assume  $x \geq y$ . We partition the committee nodes in  $m(A)$  into  $y$  classes<sup>3</sup> of equal size, and map each committee node in  $m(B)$  to a class  $y_v$  via a one to one mapping. The partitioning and mapping are predetermined before the algorithm is run. Each processor  $v \in m(B)$  will poll

---

<sup>3</sup>If  $y > x$ , we partition the committee nodes in  $|m(B)|$  into  $x$  classes and map each committee node in  $m(A)$  to a class.

## Chapter 1. Scalable Leader Election

every processor in  $y_v$ , where  $y_v$  is the class  $v$ 's committee node was mapped to.

Thus every elected processor  $u \in A$  determines the identity of the subcommittee elected in  $B$  by the following protocol, which we call  $V(A, B)$ .

### **V(A,B)**

1. Each processor  $v \in m(B)$  validates the subcommittee elected in  $A$ , by first polling  $m(A)$  according to the predetermined mapping described above, and second taking the majority to determine the subcommittee elected in  $A$ .
2. Every elected processor  $u \in A$  randomly selects a set of  $C_2 \ln n$  processors in  $m(B)$  to poll and takes the majority to determine the subcommittee elected in  $B$ .

In the course of the algorithm it will also be the case, every processor in committee node  $A$ , will need to know the identity of every processor in committee node  $B$ , again we represent this by the pair  $(A, B)$ . Let  $C(A)$  and  $C(B)$  represent the children of committee nodes  $A$  and  $B$ , respectively, in the layered network. Every processor  $u \in A$  determines the identity of every processor in  $B$  by the following protocol, which we call  $P(A, B)$ .

### **P(A,B)**

For all pairs  $(J,K)$ , where  $J \in C(A)$  and  $K \in C(B)$  do the following:

1. Each processor  $v \in m(K)$  validates the subcommittee elected in  $J$ , by first polling  $m(J)$  according to the predetermined mapping described above, and second taking the majority to determine the subcommittee elected in  $J$ .
2. Every elected processor  $u \in J$  randomly selects a set of  $C_2 \ln n$  processors in  $m(K)$  to poll and takes the majority to determine the subcommittee elected in  $K$ .

*Note in this section and throughout the chapter we implicitly assume processors ignore requests from other processors which have not been validated.*

### 1.5.3 Downward communication tree

We construct a rooted  $\ln n$ -ary tree  $T$  whose node set  $N$  is the set of committee nodes in the layered network. The tree is rooted at the single committee node in level  $i^*$ . If there are  $z$  committee nodes at level  $k$  in the tree, we assign children to these nodes by partitioning the committee nodes of level  $k - 1$  into  $z$  classes and arbitrarily mapping one class to each level  $k$  committee node. For any committee node  $A \in T$ , we let  $T(A)$  denote its children in  $T$ .

*To avoid confusion, we use the term level, when we are discussing the tree  $T$ , and the term layer, when we are discussing the layered network.*

### 1.5.4 The Communications Protocol

Assume the most recent elections have taken place on layer  $i$  in the layered network. If a committee node  $J \in T$  is a descendant of a level  $i$  committee node  $A \in T$ , we say  $J$ 's processors are responsible for communicating  $A$ 's election result. Initially the processors at level  $i$  in  $T$ , are responsible for communicating their committee node's election result. The election results of layer  $i$  are communicated down the tree  $T$ , to the monitoring sets by the following protocol, which we call  $E(i)$ .

$E(i)$

1.  $k \leftarrow i$
2. *while*{ $k > 0$ }

## Chapter 1. Scalable Leader Election

- (a) For every committee node  $A$  at level  $k$  in  $T$ , and all  $J \in T(A)$ , do  $P(A, J)$  and  $P(J, A)$ .
- (b) The processors in every committee node  $A$  at level  $k$ , communicate the layer  $i$  election results they are responsible for to every processor in a child of  $A$ .
- (c) Each processor in a child of  $A$  takes the majority to determine the election result.
- (d)  $k \leftarrow k - 1$

*Note step 2a in the while loop will have been executed previously for all  $k < i$ , when the results of previous layers are communicated down the tree. Thus in a practical implementation, step 2a would only be executed for  $k = i$ .*

## 1.6 The Leader Election Algorithm

In this section we detail the Leader Election Algorithm, which we call *Leader*.

### Leader

1.  $k \leftarrow 0$
2. *while*{ $k \leq i^*$ }
  - (a) For all committee nodes  $A$  on layer  $k$ , each processor  $v \in A$  determines the other members of  $A$  by doing the following. For all pairs  $(J, K)$ , where  $J \neq K$  and  $J, K \in C(A)$ , do  $V(J, K)$  and  $V(K, J)$ . *Note, if  $k = 0$ , the members of the committee nodes are predetermined.*
  - (b) If  $k < i^*$ , each committee node  $A$  on layer  $k$ , elects a subcommittee by running the *subcommittee election protocol*. If  $k = i^*$ , the processors elect a leader using the

*heavy weight leader election protocol.* If a good processor is elected to more than one subcommittee on layer  $k$ , it stops participating in any future elections on layers  $i > k$ .

- (c) The election results are communicated to the monitoring sets by running  $E(k)$ . *Note the monitoring set for the election on layer  $i^*$  consists of all the layer 0 committee nodes.*
- (d)  $k \leftarrow k + 1$

In order for the processors to determine the leader elected in the final election on layer  $i^*$ , the  $n$  processors are partitioned into  $n/(C_1 \ln^4 n)$  classes. Each class is mapped to a layer 0 committee node via a one to one mapping (note both the partitioning and mapping are predetermined before the algorithm is run). The processors who did not participate in the final election, query all the processors in the layer 0 committee node they were mapped to, and take the majority to determine the leader.

*We note in passing, we can simulate broadcast from the elected leader on layer  $i^*$  by using the tree  $T$  and the polling step described above. Thus in order to send a bit from the elected leader to a  $1 - o(1)$  fraction of the processors, every processor (including the leader) only sends a polylogarithmic number of bits.*

## 1.7 Proof of Theorem 1

Since the number of layers in the network is  $O(\ln n / \ln \ln n)$ , and each processor by Lemma 3 participates in  $O(\ln^4 n)$  elections per layer; it is easy to see, once *Leader* has completed, no good processor sends or processes more than a polylogarithmic number of bits. To complete the proof of Theorem 1 we will prove the following lemma:

## Chapter 1. Scalable Leader Election

**Lemma 4** *Let the fraction of bad processors,  $\beta$ , be less than  $1/3$ . Let  $\alpha = 1 - \beta$  denote the fraction of good processors. W.h.p., in Leader, a  $(1 - O(1/\ln \ln n))\alpha$  fraction of the processors on layer  $i^*$  are good. In other words, the fraction of good processors on layer  $i^*$  is greater than  $2/3$ .*

The remainder of this section will be devoted to proving Lemma 4 and showing that Theorem 1 follows from it. First we introduce the following definitions.

- Call a committee node on layer  $i$  *inherently good* if at least a  $2/3 + \epsilon$  fraction of its processors are good. Else call it *inherently bad*.
- As in Lemma 2, let  $f_S$  denote the fraction of good processors in committee node  $A$ . Call the subcommittee elected in  $A$  *inherently good*, if  $f_S > 2/3$ , and the fraction of good processors in the subcommittee is greater than  $(1 - 1/\ln n)f_S$ . Else call it *inherently bad*.
- Call a monitoring set which monitors an election  $J$ , *inherently good*, if at least a  $9/10 + \epsilon$  fraction of the committee nodes in the monitoring set are both inherently good and their good processors correctly know the result of election  $J$ . Else call it *inherently bad*.
- Call a monitoring set which monitors an election  $J$ , *inherently good* with respect to a good polling processor  $v$ , if at least a  $8/15 + \epsilon$  fraction of the processors in the set satisfy all of the following three conditions. Else call it *inherently bad*.
  - The processors are good.
  - The processors correctly know the result of election  $J$ .
  - The processors are able to correctly validate the identity of  $v$ .

**Observation 2** *Let the set  $m(A)$  be an inherently good monitoring set with respect to election  $A$ . Let  $s_A$  be the subcommittee elected in  $A$ . Let the set  $m(B)$  be an inherently good monitoring set with respect to election  $B$ , where  $m(B)$  will be polled by the members of  $s_A$ . For every good processor  $v \in s_A$ ,  $m(B)$  is inherently good with respect to  $v$ .*

## Chapter 1. Scalable Leader Election

Let  $P_4$  be the probability a good processor  $v$ , correctly determines an election result by polling a monitoring set inherently good with respect to  $v$ .

**Lemma 5** *For any fixed  $c$ , the constant  $C_2$ , in the protocol  $V(A, B)$  can be chosen such that  $P_4 \geq 1 - n^{-c}$ . Hence w.h.p., every such poll succeeds during the course of Leader.*

**Proof** By definition, at least a  $8/15$  fraction of the processors in the monitoring set, are good, correctly know the election result, and will respond to  $v$ . The lemma now directly follows from an application of the Chernoff bound and Boole's inequality.  $\square$

For the remainder of the proof we treat every processor in an inherently bad subcommittee as being bad. Additionally, if an elected subcommittee is *not* monitored by an inherently good monitoring set, we will treat the subcommittee as being inherently bad. Also, if a *good* processor is elected to more than one subcommittee on layer  $i$ , in all layers  $k > i$ , we will treat the processor as being bad<sup>4</sup>.

**Lemma 6** *If all the good processors in an inherently good committee node  $J$  correctly know the result of election  $A$ , w.h.p, for each inherently good committee node  $K \in T(J)$ , all the good processors in  $K$  correctly know the result of election  $A$ .*

**Proof** This follows from the definition of an inherently good committee and Lemma 5.  $\square$

**Lemma 7** *Let  $z$  be the number of committee nodes on layer  $i$ . Note  $z$  is also the number of monitoring sets which monitor a layer  $i$  election. If for every layer  $k \leq i$ , a  $1 - O(1/\ln^2 n)$  fraction of committee nodes on layer  $k$  are inherently good, w.h.p, the number of monitoring sets for the layer  $i$  elections which are inherently bad, is  $O(z/\ln n)$ .*

**Proof** Since the number of layers in the network is  $O(\ln n / \ln \ln n)$ , by the assumptions of the lemma and repeated application of Lemma 6, w.h.p., a  $1 - O(1/\ln n)$  fraction of the layer 0

---

<sup>4</sup>Since the processor does not participate in any elections on layers  $k > i$ .

committee nodes are both inherently good and correctly know the result of the election monitored by their monitoring set. Since, for the monitoring set to be inherently bad, at least a  $1/10$  fraction of the committee nodes in any monitoring set must *not* satisfy the above criteria, the lemma follows.  $\square$

**Lemma 8** *Assume a  $1 - O(1/\ln^2 n)$  fraction of committee nodes on layer  $i$  are inherently good. Let  $f_L$  be the fraction of good processors on layer  $i$ . W.h.p., a  $O(f_L/\ln n)$  fraction of the good processors elected on layer  $i$  become bad by assuming processors elected to multiple subcommittees on the same layer are bad.*

**Proof** Let  $X$  be the total number of processors participating in elections on layer  $i$ . The number of processors elected on layer  $i$ , counting multiplicity, is  $X/\ln n$ . We show that w.h.p., the number of good processors elected more than once, counting multiplicity, is  $O(X/\ln^2 n)$ .

By construction, the number of elections on layer  $i$  is  $X/(C_1 \ln^4 n)$ , where each election employs  $\ln^5 n$  bins, and no processor participates in more than  $2 \ln^4 n$  elections. Assume for each election the adversary is able to choose which bin is elected. For a fixed sequence of  $X/(C_1 \ln^4 n)$  bin choices, let  $p_k$  be the probability that a given (good) processor is elected  $k$  or more times. Since the good processors make their bin choices independent of each other, for a fixed bin sequence, the probability that at least  $\delta X$  processors are elected  $k$  or more times is

$$\leq \binom{X}{\delta X} p_k^{\delta X} \leq \left(\frac{e p_k}{\delta}\right)^{\delta X}.$$

Since the total number of possible bin sequences is at most  $(\ln^5 n)^{X/\ln^4 n}$ , the probability that any bin sequence causes at least  $\delta X$  processors to be elected  $k$  or more times is

$$\leq \left(\frac{e p_k}{\delta}\right)^{\delta X} \cdot (\ln^5 n)^{X/\ln^4 n}. \tag{1.1}$$

Recalling that no processor participates in more than  $2 \ln^4 n$  elections,

$$p_k \leq \binom{2 \ln^4 n}{k} \cdot \left(\frac{1}{\ln^5 n}\right)^k \leq \left(\frac{2e}{k \ln n}\right)^k.$$

## Chapter 1. Scalable Leader Election

For  $2 \leq k \leq 16$ , we set  $\delta = 18^k / \ln^2 n$ , for  $16 < k \leq 2 \ln^4 n$ , we set  $\delta = 12 / (k \ln^4 n)$ . For these choices of  $\delta$ , since  $X \geq \ln^{10} n$ , we see from (1.1) that the probability that more than  $\delta X$  processors are elected  $k$  or more times is  $o(n^{-3})$ . Hence, w.h.p., the number of good processors elected more than once, counting multiplicities, is at most

$$2 \cdot \left( \sum_{k=2}^{16} 18^k / \ln^2 n + \sum_{k=17}^{2 \ln^4 n} 12 / (k \ln^4 n) \right) = O(X / \ln^2 n).$$

Since a  $1 - O(1 / \ln^2 n)$  fraction of committee nodes on layer  $i$  are inherently good, by Lemma 2 the number of good processors elected on layer  $i$ , counting multiplicity, is  $(f_L X / \ln n)(1 - O(1 / \ln n))$ . Thus the lemma follows.  $\square$

**Lemma 9** *Assume a  $1 - O(1 / \ln^2 n)$  fraction of committee nodes on all layers  $k \leq i$  are inherently good. As in the previous lemma let  $f_L$  be the fraction of good processors on layer  $i$ . W.h.p., the fraction of good processors elected on layer  $i$ , is  $(1 - O(1 / \ln n))f_L$ .*

**Proof** By Lemma 3 in a  $1 - O(1 / \ln^2 n)$  fraction of the committees nodes, the fraction of good processors is  $(1 - O(1 / \ln n))f_L$ . By Lemma 2, w.h.p., no inherently good committee node elects an inherently bad subcommittee. By Lemma 7, w.h.p., a  $O(1 / \ln n)$  fraction of the elections on layer  $i$  are monitored by inherently bad monitoring sets. By Lemma 8, w.h.p, a  $O(f_L / \ln n)$  fraction of the good elected processors become bad by treating processors which win multiple elections on the same layer as bad. The lemma now readily follows.  $\square$

**Lemma 10** *Assume  $\beta < 1/3$ . As in the previous lemmas, let  $f_L$  be the fraction of good processors on layer  $i$ . For every layer  $i$ , w.h.p., the fraction of good processors elected on layer  $i$ , is  $(1 - O(1 / \ln n))f_L$ .*

**Proof** We prove the lemma inductively. Since  $\beta < 1/3$ , by Lemma 3, initially a  $1 - O(1 / \ln^2 n)$  fraction of the layer 0 committee nodes are inherently good, and the fraction of good processors in those committee nodes is  $(1 - O(1 / \ln n))\alpha$ . Thus Lemma 9 holds. Hence for layer 0,

## Chapter 1. Scalable Leader Election

Lemma 10 is true. Assume Lemma 10 is true for all layers  $k \leq i$ . Since the number of layers is  $O(\ln n / \ln \ln n)$ ; by the inductive hypothesis, and Lemmas 3 and 5, w.h.p., for all layers  $k \leq i + 1$ , a  $1 - O(1/\ln^2 n)$  fraction of layer  $k$ 's committee nodes are inherently good, and the fraction of good processors in those committee nodes is  $(1 - O(1/\ln n))^i \alpha$ . Thus Lemma 9 holds for layer  $i + 1$ . Hence Lemma 10 holds for layer  $i + 1$ .  $\square$

Since the number of levels is  $O(\ln n / \ln \ln n)$ , Lemma 4 follows by the repeated application of Lemma 10.

We now complete the proof of Theorem 1. By Lemmas 3, 5, and 10, w.h.p, a  $1 - O(1/\ln^2 n)$  fraction of the committee nodes on every layer  $k < i^*$  are inherently good. Also by Lemma 4, w.h.p., the committee node on layer  $i^*$  is inherently good. Thus by repeated application of Lemma 6, w.h.p, a  $1 - O(1/\ln n)$  fraction of the layer 0 committee nodes are both inherently good and know the leader elected on layer  $i^*$ . Since in the final polling step (after the while loop in *Leader* has terminated) the  $n$  processors are partitioned among the layer 0 committee nodes, w.h.p., a  $1 - O(1/\ln n)$  fraction of the good processors know the leader elected on layer  $i^*$ . Finally, by Lemmas 1 and 4, with constant probability a good leader is elected on layer  $i^*$ . Theorem 1 now follows by taking a union bound.

## **Chapter 2**

# **The chromatic number of random scaled sector graphs**

### **2.1 Acknowledgements**

I would like to thank Josep Diaz, Maria Serna, and Paul Spirakis. The work in this chapter was done jointly with them, a preliminary version of which appeared in [25].

### **2.2 Introduction**

Massive networks of wireless sensors are known to play an important role in monitoring and disseminating information [21, 20]. The general setting of such a network is to have a large collection of wireless motes (sensors) randomly scattered in a remote or hazardous terrain, performing tasks of distributed sensing. The sensing information gathered by the motes is relayed to a base station. To communicate, either among themselves or with a monitoring base station, the motes use radio-frequency (RF) or optical communication. In the RF communication model, the

motes either use an omnidirectional antenna, which spreads the signal in a spherical region centered at the antenna, or a directional antenna, which has a focused beam spanning a sector of  $\alpha$  degrees. In sensor networks, directional antennas have multiple advantages over omnidirectional antennas: less energy consumption, less fading area, and furthermore as the transmission area is smaller the channel interference may have less influence [22]. In the optical communication model motes can send information using an orientable laser beam embedded with an optical receiver. In this model motes can receive information from any mote within a prescribed distance whose transmitting laser is orientated towards them [26].

In recent times, there has been an effort to provide a theoretical framework to study networks of sensors. For the omnidirectional RF communication network, a suitable model is the *random geometric graph*, also denoted *random scaled disk graph*. These graphs are the random scaled version of the *unit disk graphs* described in [23]. The model considers the network as a graph scaled into  $[0, 1]^2$ , where the  $n$  random deployed motes are the vertices of a random graph in  $[0, 1]^2$ , and two vertices are connected if they are within Euclidean distance  $r_n$ , corresponding to the broadcast range of the motes. Many results are known about the properties of random scaled disk graphs. For instance, when  $r_n = \sqrt{\frac{\ln n}{n}}$ , it is known the chromatic number  $\chi$  and the clique number  $\omega$  are asymptotically  $\Theta(\ln n)$  (see [29]).

The natural model for the case of directional RF and optical networks seems to be the *random scaled sector graph*, a generalization of the random geometric graph, introduced in [24]. In the setting under consideration, each vertex  $i$  is assigned a uniformly at random sector  $\beta_i$ , of fixed central angle  $\alpha$  ( $0 < \alpha \leq 2\pi$ ), defining a sector of transmission. We represent the beam emitted by  $i$  as the sector  $S_i$ , centered at  $i$ , with radius  $r$ , amplitude  $\alpha$  and orientation angle  $\beta_i$ . Every other sensor which falls inside of  $S_i$  can potentially receive the signal emitted by  $i$  (see Fig. 1). The random scaled sector graph is the graph with vertices as the sensors, in which there is an arc from  $i$  to  $j$  if  $j$  falls inside  $S_i$ , (see formal definition in Section 2). Some of the graph parameters for sector graphs coincide with the ones for geometric graphs, for instance in both graphs the threshold for connectivity, in terms of the distance  $r$  is  $r_n = \Theta\left(\sqrt{\frac{\ln n}{n}}\right)$  (see [24]). It should

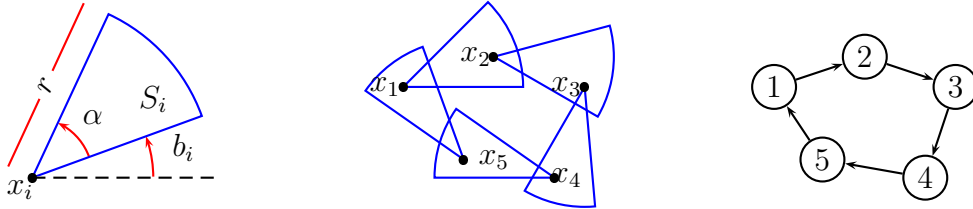


Figure 2.1: The sector of a sensor  $i$  and the communication between motes

be noted, that in *practical applications*, the values of  $\alpha$  are small, typically from  $\pi/20$  to  $\pi/4$ , depending on the type of communication (RF or optical).

In this chapter we study the value of the chromatic number  $\chi(G_n)$ , directed clique number  $\omega(G_n)$ , and undirected clique number  $\widehat{\omega}_2(G_n)$  for random scaled sector graphs with  $n$  vertices and radius  $r_n = \sqrt{\frac{\ln n}{n}}$ . Asymptotically, we prove that for values  $\alpha < \pi$ , as  $n \rightarrow \infty$  w.h.p.,  $\chi(G_n)$  is  $\Theta\left(\frac{\ln n}{\ln \ln n}\right)$ , showing a clear difference with the random geometric graph model. For  $\alpha > \pi$ , w.h.p. the value of  $\chi(G_n)$  is  $\Theta(\ln n)$  for both random sector and geometric graphs.

## 2.3 Results

A random scaled sector graph is defined in the following way,

**Definition 1 ([24])** Assume that  $\alpha$  is a fixed parameter of the sensors. Let  $X = (x_i)_{i \geq 1}$  be a sequence of independently and uniformly distributed (i.u.d.) random points in  $[0, 1]^2$ , let  $B = (\beta_i)_{i \geq 1}$  be a sequence of i.u.d. angles and let  $R = (r_i)_{i \geq 1}$  be a sequence of numbers in  $[0, 1]$ . We write  $X_n = \{x_1, \dots, x_n\}$  and  $B_n = \{\beta_1, \dots, \beta_n\}$ . We call the digraphs  $G_n = \mathcal{G}_\alpha(X_n, B_n, r_n)$  the random scaled sector graph on  $n$  nodes, where  $V(G_n) = X_n$  and the arcs are defined by:

$(x_i, x_j) \in E(G_n)$  iff  $x_j \in S_i$ .

We use the letter  $H$  to denote a subgraph of  $G_n$ .  $\Delta$ , denotes the maximum degree of  $G_n$ . Given  $G_n$ , as usual the chromatic number, and the size of the maximum directed clique, are represented by  $\chi(G_n)$  and  $\omega(G_n)$ , respectively. Since we are dealing with directed graphs, we introduce a new variable  $\widehat{\omega}_2$ , which represents the size of the maximum undirected clique, where for any two vertices  $u, v \in V(G_n)$ , to be members of the same undirected clique, only one of the two possible arcs,  $(u, v)$  or  $(v, u)$ , need be present in the graph  $G_n$ . Thus,  $\omega(G_n) \leq \widehat{\omega}_2(G_n) \leq \chi(G_n)$ , and for  $\alpha = 2\pi$ ,  $\omega(G_n) = \widehat{\omega}_2(G_n)$ .

We say  $G_n$  has a property T, with high probability (w.h.p), if as  $n \rightarrow \infty$ , we expect  $G_n$  to have property T, with probability  $1 - O(1/n^c)$ , for some  $c > 0$ . For other concepts and results in probability theory, look for example [29].

In the remainder of the chapter we prove the following results for  $r_n = \sqrt{\frac{\ln n}{n}}$ ,

**Theorem 1** *Let  $\epsilon > 0$ . For  $\epsilon < \alpha < \pi - \epsilon$ , the size of the maximum directed clique,  $\omega(G_n)$  is  $\Theta(1)$ . For  $\pi + \epsilon < \alpha < 2\pi - \epsilon$ , w.h.p.,  $\omega(G_n)$  is  $\Theta\left(\frac{\ln n}{\ln \ln n}\right)$ .*

**Theorem 2** *Let  $\epsilon > 0$ . For  $\epsilon < \alpha < \pi - \epsilon$ , w.h.p., the chromatic number,  $\chi(G_n)$  is  $\Theta\left(\frac{\ln n}{\ln \ln n}\right)$ . For  $\pi + \epsilon < \alpha$ , w.h.p.,  $\chi(G_n)$  is  $\Theta(\ln n)$ .*

**Theorem 3** *Let  $\epsilon > 0$ . For  $\epsilon < \alpha < \pi - \epsilon$ , w.h.p., the size of the maximum undirected clique,  $\widehat{\omega}_2(G_n)$  is  $\Theta\left(\frac{\ln n}{\ln \ln n}\right)$ . For  $\pi + \epsilon < \alpha$ , w.h.p.,  $\widehat{\omega}_2(G_n)$  is  $\Theta(\ln n)$ .*

## 2.4 Basic constructions and lemmas

In this section, we present some tools and lemmas, which are needed to prove Theorems 1, 2, and 3. In order to lighten the notation we define the following variables,

$$a_n = \frac{\ln n}{\ln \ln n} \quad \text{and} \quad b_n = \sqrt{n \ln n}.$$

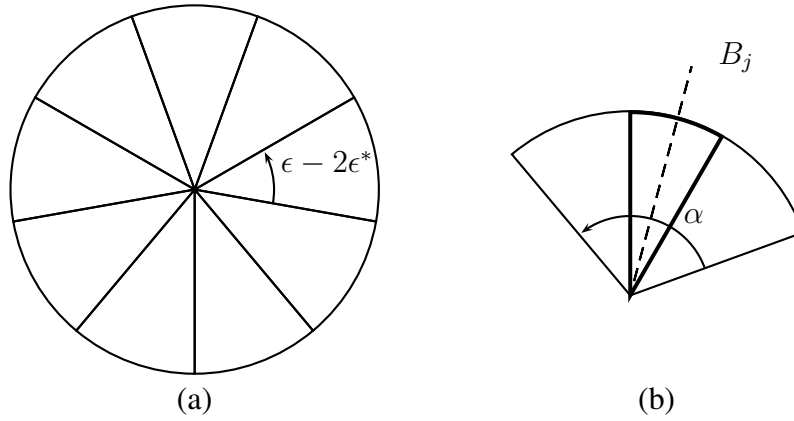


Figure 2.2: Angle partition for  $\alpha > \pi + \epsilon$  (a) classes  $\mathcal{B}$  (b) directions associated to a class  $B_j$

Recall, the orientation angle,  $\beta_i$ , of every mote  $i$  is drawn uniformly at random (u.a.r) from  $(0, 2\pi]$ . Many of the proofs in this chapter require partitioning the orientation angle into classes. Thus we define a partition  $\mathcal{B}$  of the orientation angle as follows.

**Definition 2** Let  $\epsilon$  be a constant (depending on  $\alpha$ ), such that  $\alpha = \pi + \epsilon$ , for  $0 < \epsilon < \pi$ . A  $\mathcal{B}$  partition, is a partition of the region  $2\pi$  into  $B$  classes, each of length  $\epsilon - 2\epsilon^*$ , with  $\epsilon^*$  a constant chosen such that  $\epsilon > 2\epsilon^*$  (see Figure 2.2). All motes  $i$  such that  $\beta_i$  fall whose the same range will belong to the same class. More specifically, for any  $1 < j \leq B$ , the class  $B_j$  is defined as the class of motes whose bisectrix falls between  $(-\frac{3}{2} + j)\epsilon - (2j - 3)\epsilon^*$  and  $(-\frac{1}{2} + j)\epsilon - (2j - 1)\epsilon^*$ . Notice  $B = \lceil \frac{2\pi}{\epsilon - 2\epsilon^*} \rceil$ , so  $B \in \mathbb{Z}$ .

Throughout this chapter, when we refer to the *dissection*  $\mathcal{S}$  of  $[0, 1]^2$ , we mean a partition of  $[0, 1]^2$  into  $\frac{n}{\ln n}$  squares, each one of size  $r_n \times r_n$ . Also, in this chapter we make use of the following three lemmas.

**Lemma 1 ([24])** If  $n$  motes are distributed u.a.r. on  $[0, 1]^2$ , w.h.p. each of the squares in the dissection  $\mathcal{S}$  will contain  $\Theta(\ln n)$  motes.

**Lemma 2** If  $n$  motes are distributed u.a.r. on  $[0, 1]^2$ , w.h.p. every square in the dissection  $\mathcal{S}$  will contain at most  $3 \ln n$  motes.

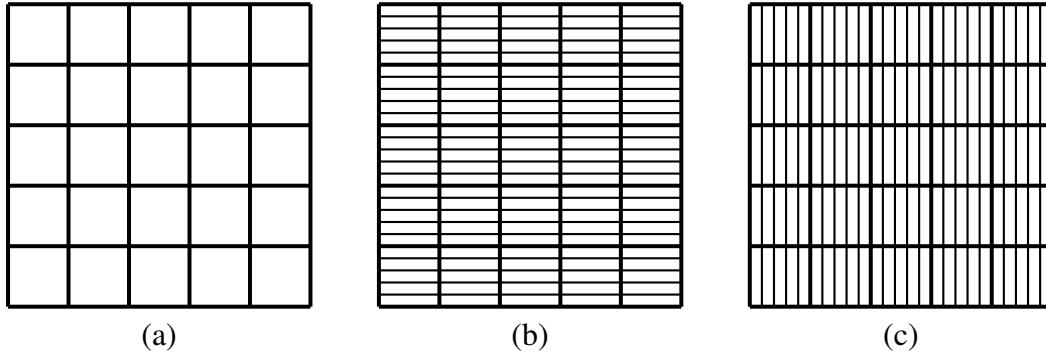


Figure 2.3: The basic dissections of  $[0, 1]^2$  (a)  $\mathcal{S}$  (b) horizontal subdivision (c) vertical subdivision

**Lemma 3** *Given the dissection  $\mathcal{S}$  of  $[0, 1]^2$ , divide each square of  $\mathcal{S}$  into  $\ln n$  rectangular regions of size  $\frac{r_n}{\ln n} \times r_n$  (see Figure 2.3). Then, w.h.p. there exists at least one region  $R_i$ , which contains  $(1 - o(1))\frac{a_n}{B}$  motes from every class  $B_j \in \mathcal{B}$ .*

The first two lemmas can be established via Chernoff bounds and Boole's inequality. In order to prove the third lemma, we use an implication of Talagrand's inequality, given in [28]:

**Talagrand's Inequality** *Let  $X$  be a non-negative random variable, not identically 0, which is determined by  $n$  independent trials  $L_1, \dots, L_m$ , and satisfying the following for some  $b, r > 0$ :*

1. Changing the outcome of any one trial can affect  $X$  by at most  $b$ ,
2. for any  $s$ , if  $X \geq s$  there is a set of  $< rs$  trials whose outcomes certify that  $X \geq s$ .

*Then, for any  $0 \leq l \leq \mathbf{E}[X]$ ,  $\mathbf{P}\left(|X - \mathbf{E}[X]| > l + 60b\sqrt{r\mathbf{E}[X]}\right) \leq 4e^{-l^2/8b^2r\mathbf{E}[X]}$ .*

**Proof of Lemma 3** In order to prove Lemma 3 we make use of the following fact. If  $n$  balls are dropped into  $n$  bins, w.h.p. at least one bin contains  $a_n = \ln n / \ln \ln n$  balls. Notice by construction the number of regions in  $[0, 1]^2$  is  $\frac{n}{\ln n} \ln n = n$ , since the  $n$  motes are distributed u.a.r. on  $[0, 1]^2$ , by a *balls-and-bins* argument, there is a region  $R_i$ , which w.h.p. contains  $a_n = \ln n / \ln \ln n$  motes. Let  $X_j$  be a random variable counting the number of motes in  $R_i$

which are in class  $B_j$ . Then  $\mathbf{E}[X_j] = a_n/B$ . To complete the proof of Lemma 3 we show via Talagrand's inequality the random variable  $X_j$  is concentrated around its expectation. First note,  $X_j$  is determined by the  $m = (1 - o(1))a_n/B$  trials specifying  $\{\beta_1, \dots, \beta_m\}$ . Also changing the outcome of any one  $\beta_l$ ,  $1 \leq l \leq m$ , can affect  $X_j$  by at most one, and in order to certify  $X_j \geq s$ , only the outcomes of  $s$  trials (the  $s$   $\beta_l$ 's which fall in that class) are required. Thus the conditions of Talagrand's inequality are satisfied with  $b, r = 1$ . Hence by Talagrand's inequality and Boole's inequality, w.h.p. every class contains  $(1 - o(1))a_n/B$  motes.  $\square$

## 2.5 Proof of Theorem 1

### 2.5.1 $\alpha < \pi - \epsilon$

**Proof** When  $\alpha < \pi - \epsilon$  the vertices of any clique must form a convex polygon. This can be proved by first noting that in every clique of size three, the three points cannot be collinear, and proceeding inductively. Let  $|V|$  represent the number of vertices in any clique. Since the sum of the angles of a convex polygon is  $(|V| - 2)\pi$ , we have  $|V|\alpha \geq (|V| - 2)\pi$ , thus  $\omega(G_n) \leq \lfloor \frac{2\pi}{\pi - \alpha} \rfloor$ .  $\square$

### 2.5.2 $\alpha > \pi + \epsilon$

**Proof** • First we establish the lower bound, by proving a certain sufficient configuration of motes exists (w.h.p.). Consider the  $\mathcal{S}$  partition of  $[0, 1]^2$ . Subdivide each small square into  $\ln n$  equal (in terms of area) vertical regions (one can imagine drawing  $\ln n$  equally spaced vertical lines). By Lemma 3, there is a vertical region  $R_i$  w.h.p. containing  $(1 - o(1))\frac{a_n}{B}$  motes who are members of the class  $B_1$ , i.e. the bisectrix of these motes is between  $-\frac{1}{2}\epsilon + \epsilon^*$  and  $\frac{1}{2}\epsilon - \epsilon^*$ . Let  $M_1$  be the set of these motes (in class  $B_1$ ) in  $R_i$ . Further subdivide the region  $R_i$  into  $a_n/B$  cells, each cell a rectangle of width  $\frac{1}{b_n}$  and height  $\frac{B r_n}{a_n}$ , see Fig. 2.4. Let  $Y$  be a random variable

counting the number of cells containing at least one vertex from  $M_1$ , then  $\mathbf{E}[Y] = (1 - \frac{1}{e})a_n/B$  as  $n \rightarrow \infty$ , and as in the proof of Lemma 3, one can show  $Y$  is concentrated around its expectation by applying Talagrand's inequality. Thus w.h.p there are at least  $(1 - (\frac{1}{e} + o(1)))a_n/B$  cells containing at least one mote from  $M_1$ . Consider a mote  $m$  from the set  $M_1$ , due to its orientation angle, it will have an arc with any mote  $m'$  which is an cell more than a specific distance,  $l$ , and within distance  $r_n$  in either direction, up or down from itself. Where  $l$  depends on the exact orientation angle of the mote  $m$ , and the location of the two motes,  $m$  and  $m'$  in their respective cells. Consider the worst case, assume the mote  $m$  is in the lower right-hand corner of its cell and the mote  $m'$  is in the upper left-hand corner of its cell (see Fig. 2.4). In this case assuming  $m$  has a bisectrix of 0, by trigonometry,  $l = \left\lceil \frac{\cos((\alpha-\pi)/2)}{\sin((\alpha-\pi)/2)b_n} \right\rceil$ . However  $m$  need not have a bisectrix of 0. Since  $m$  is in the class  $B_1$ , its bisectrix is between  $-\frac{1}{2}\epsilon + \epsilon^*$  and  $\frac{1}{2}\epsilon - \epsilon^*$ , thus in the worst case,  $l = \left\lceil \frac{\cos((\alpha-\pi-\epsilon+2\epsilon^*)/2)}{\sin((\alpha-\pi-\epsilon+2\epsilon^*)/2)b_n} \right\rceil$ . Recall  $\alpha > \pi + \epsilon$ , thus when  $\alpha$  assumes its lowest value,  $l = \left\lceil \frac{\cos(\epsilon^*)}{\sin(\epsilon^*)b_n} \right\rceil$ . For small  $x$ ,  $\sin(x) \sim x$ , given that  $\epsilon^*$  is a constant,  $l = c/b_n$ , for some constant  $c$ . Since the height of each cell is  $\frac{Br_n}{a_n}$ , w.h.p.,  $\omega(G_n) \geq c'a_n$ , for a sufficiently large constant  $c'$  dependent on  $\alpha$ .

- Next we establish the upper bound, by showing w.h.p., a certain necessary configuration cannot exist. In order to prove the upper bound we make use of the following easily verified fact, let  $\omega^*$  represent the size of the largest directed clique in any square of  $\mathcal{S}$ , then the size of  $\omega$  is upper-bounded by  $9\omega^*$ . Thus to establish the upper bound, we will prove w.h.p., there exists a sufficiently large constant  $d$  such that no set of  $\frac{d}{B}a_n$  motes in any square  $\mathcal{S}$  form a clique, i.e.  $\omega^* \leq \Theta(a_n)$  and the statement of the theorem will follow.

Again consider the partition  $\mathcal{S}$  of  $[0, 1]^2$  and the  $\mathcal{B}$ -partition. Fix any square  $S \in \mathcal{S}$ . By Lemma 2,  $S$  contains at most  $3 \ln n$  motes. Select u.a.r.  $d \cdot a_n$  motes from  $S$ . By the Pigeon-hole principle, at least  $\frac{d}{B}a_n$  of those motes will have the bisectrix oriented into the same class, call this class  $B_j \in \mathcal{B}$ . Let  $M_j$  be the set of all those motes in the class  $B_j$ . Define a partition of  $S$  into  $\ln n$  strips in the following way: Imagine a mote exists in  $S$  whose bisectrix is orientated exactly in the center of the class  $B_j$ . Draw a parallel line to the bisectrix of this (imaginary) mote. This

(parallel) line will be the *orientation* of the strips. Cover  $S$  with  $\ln n$  rectangular strips parallel to the orientation (see Figure 2.5). For example, in the case where these motes belong to the class  $B_1$ , the rectangular strips will be parallel to the sides of  $S$ . Note by construction, in this partition of  $S$ , the optical sensors of all the motes in the class  $M_j$  look in the same approximate direction. Thus for these motes to be a part of the same clique every mote must see all the other motes along some specified direction, and we will show w.h.p. this will not be the case.

Before we continue with the remainder of the proof we need to establish the following lemma.

**Lemma 4** *For a sufficiently large but constant  $d$ , any set of  $\frac{d}{B}a_n$  motes, will (w.h.p.) occupy at least  $(\ln n)^{9/10}$  strips.*

**Proof** First we upperbound the area of any strip. The maximum length any strip can have is  $r_n\sqrt{2}$  since we are considering a  $[r_n \times r_n]$  square  $S$ . The maximum width any strip can have is  $\sqrt{2}r_n/\ln n$  (this occurs when the orientation of strips is parallel to the diagonal of  $S$ ). Thus the area of the largest strip is bounded above by

$$\sqrt{2}r_n/\ln n \times r_n\sqrt{2} = \frac{2}{n} .$$

Now we upperbound the probability that any set of  $(\ln n)^{9/10}$  of the strips will contain  $\frac{d}{B}a_n$  motes in  $S$ .

Any set of  $(\ln n)^{9/10}$  strips by the above upperbound have total area at most  $\frac{2\ln n^{9/10}}{n}$ . Thus the area of any set of  $(\ln n)^{9/10}$  strips divided by the area of  $S$  is at most  $\frac{2}{(\ln n)^{1/10}}$ , which is the probability that any given mote in  $S$  falls in the  $(\ln n)^{9/10}$  strips.

Let  $p_1$  be the probability that in any small square, a set of least  $\frac{d}{B}a_n$  motes falls in at most  $(\ln n)^{9/10}$  strips. W.h.p. no small square has more than  $3 \ln n$  motes, thus the number of ways to choose a set of  $\frac{d}{B}a_n$  motes from  $3 \ln n$  motes is

$$\binom{3 \ln n}{\frac{d}{B}a_n} < n^3 .$$

Chapter 2. The chromatic number of random scaled sector graphs

Moreover, as there are  $n/\ln n$  small squares and at most  $n$  ways to choose  $(\ln n)^{9/10}$  strips out of  $\ln n$  strips, by Boole's inequality,

$$p_1 \leq n^5 \left( \frac{2}{(\ln n)^{1/10}} \right)^{\frac{d \ln n}{B \ln \ln n}} \leq n^6 \left( \frac{1}{e^{\frac{\ln(\ln n) d a_n}{10} - \frac{d}{B}}} \right) = n^6 e^{-\frac{d \ln n}{10B}}.$$

Therefore, as  $n \rightarrow \infty$ , a sufficiently large constant  $d$  can be chosen such that  $p_1 \rightarrow 0$ .

□

Given the above partition of  $S$  in  $\ln n$  strips, we ignore the first  $\sqrt{\ln n}$  and last  $\sqrt{\ln n}$  strips (keeping the middle strips of larger area). Every strip by construction will either have height or width  $\Theta(1/b_n)$ . Without loss of generality assume the orientation of the strips is such that the width is  $\Theta(1/b_n)$ . Define the *average height* of a strip as the average of the two sides of the strip. Consider the worst case, when the difference in height between both sides of a strip is maximal, i.e. the case where the orientation of the partition is either  $\pi/4$  or  $3\pi/4$ . Notice that the average height of all middle strips will be larger than the average height of any of the first  $\sqrt{\ln n}$  strips (strip  $T_i$  in Figure 2.5). Draw a diagonal line  $L$  of length  $\sqrt{\frac{n}{\ln n}}$ ,  $L$  spans  $\sqrt{\ln n}$  of the discarded strips. The triangle with sides  $L$ ,  $L'/2$  and the edge of  $S$  is rectangle with two angles of  $\pi/4$ , so  $L' = \Theta(\sqrt{\frac{\ln n}{n \ln n}}) = \Theta(\frac{1}{\sqrt{n}})$ . In the same way, considering the triangle formed by  $L + \Theta(1/b_n)$  and  $L''/2 = \Theta(\frac{1}{2\sqrt{n}})$  together with the side of  $S$ , the average height of strip  $T_i$  is  $\Theta(\frac{1}{\sqrt{n}})$ , and the area of any middle strip is at least the area of  $T_i$ , which is  $\Theta(\frac{1}{\sqrt{n \ln n}} \times \frac{1}{\sqrt{n}}) = \Theta(\frac{1}{n \sqrt{\ln n}})$ .

Using the same arguments used in the proof of lemma 4, we can find a sufficiently large constant  $d$  such that w.h.p., at least  $\frac{d}{2B} a_n$  motes will fall outside of the first and last  $\sqrt{\ln n}$  strips. Consider only these  $\frac{d}{2B} a_n$  motes and label the motes along the specified direction, in the following way: Scan an imaginary line along the orientation of the strips through the  $\ln - 2\sqrt{\ln n}$  strips. Label the motes from  $m_1$  to  $m_{\frac{d}{2B} a_n}$ , according to the order they are scanned by the line. Partition the motes into disjoint pairs of consecutive motes; motes  $m_{2i-1}$  and  $m_{2i}$  form a pair. For each pair of motes, each of the two motes could be in the same strip or in different strips. Since we have  $\Theta(\ln n)$  strips and  $\frac{d}{4B} a_n$  mote pairs, by the pigeon hole principle, going along the specified direction, for  $d$  chosen sufficiently large in at least  $\frac{d}{8B} a_n$  motes pairs, the two motes

in the pair, will be within  $2 \ln \ln n$  strips of each other. Now in order for these motes to be part of the same clique, in each pair both motes must see each other. Without loss of generality assume the orientation of the strips is parallel to the side of  $S$ , such that every mote can see every other mote to its *right*. Thus at least one of the necessary arcs is present. For the other arc to be present the right-most mote (in the mote pair) must see the mote to its left. Since the strips in question have a width of at least  $\Theta\left(\frac{1}{\sqrt{n}}\right)$ , the horizontal coordinates of both points are drawn u.a.r. from  $\left(0, \Theta\left(\frac{1}{\sqrt{n}}\right)\right]$ . Thus in order to compute the probability of this event we will consider two disjoint cases.

**Case one**, the horizontal coordinates of at least one mote is in the interval  $\left(0, \frac{1}{(\ln n)^{1/10}\sqrt{n}}\right]$ . In that case we will assume with probability one, the right mote see's the left mote. The probability of case one occurring is  $\Theta\left(\frac{1}{(\ln n)^{1/10}}\right)$ .

**Case two**, the horizontal coordinates of both motes is  $> \frac{1}{(\ln n)^{1/10}\sqrt{n}}$ . In this case since  $\epsilon^*$  is a constant, the maximum area a mote see's of any strip which is within  $2 \ln \ln n$  strips of it (in a specified direction) is at most  $\Theta(1/(na_n))$ . This follows, since the region of any one strip the mote see's has at most a width of  $\Theta(r_n/a_n)$  and height  $\Theta(1/b_n)$  (given the strip in question is within  $2 \ln \ln n$  strips of the mote). Now the left-most mote (in the pair) must fall in a strip. Also (since we are conditioning on being in case two), its horizontal coordinate is drawn u.a.r. from an interval of length  $> \frac{1}{(\ln n)^{1/10}\sqrt{n}}$ . Thus since every strip has height of  $\Theta(1/b_n)$ , conditioning on the particular strip the left-most mote falls into, the left-most mote falls u.a.r. into an area  $> \Theta\left(\frac{1}{n(\ln n)^{6/10}}\right)$ . Thus the probability the right mote see's the left mote, conditioned on case two occurring, is at most  $\Theta\left(\frac{1}{(\ln n)^{4/10}}\right)$ . Thus the probability the right mote see's the left mote is at most,

$$\Theta\left(\frac{1}{(\ln n)^{1/10}}\right) + \Theta\left(\frac{1}{(\ln n)^{4/10}}\right) = \Theta\left(\frac{1}{(\ln n)^{1/10}}\right) .$$

Since for every pair of motes these events are independent of each other (because only disjoint pairs are being considered), the probability for every pair of motes, both motes see each other is

$$\leq \Theta \left( \frac{1}{(\ln n)^{1/10}} \right)^{\frac{d \ln n}{8B \ln \ln n}} .$$

Let  $p_2$  denote the probability in any square  $S$  in  $\mathcal{S}$ , there is a clique of size  $d \cdot a_n$  or greater. Since (w.h.p.) in any square  $S$  we have at most  $n^3$  sets of size  $d \cdot a_n$  or greater, and  $n / \ln n$  squares in  $\mathcal{S}$ , by Boole's inequality

$$p_2 \leq n^4 \Theta \left( \frac{1}{(\ln n)^{1/10}} \right)^{\frac{d \ln n}{8B \ln \ln n}} \approx n^4 e^{-\frac{d \ln n}{80B}} .$$

Therefore, there is a sufficiently large constant  $d$  such that  $p_2 \rightarrow 0$ , and thus w.h.p.,  $\omega^* \leq \Theta(a_n)$ .

□

### 2.5.3 $\alpha = 2\pi$

For  $\alpha = 2\pi$ , a random sector graph is equivalent to a random disk graph. For a random disk graph it is already known w.h.p.,  $\omega(G_n)$  is  $\Theta(\ln n)$  [29]. However, this fact can be directly verified, as above, by partitioning the  $[0, 1]^2$  unit square into  $c \frac{n}{\ln n}$  regions, and bounding (w.h.p.) the number of motes in any region. Thus the value of  $\omega(G_n)$ , for the particular value of  $r_n$  considered in this chapter, exhibits two transitions, one at  $\pi + \epsilon$ , the other at  $2\pi$ .

## 2.6 Proof of Theorem 2

### 2.6.1 $\alpha > \pi + \epsilon$

**Proof** Partition the unit square into  $2n / \ln n$ ,  $\left[ \frac{r_n}{\sqrt{2}} \times \frac{r_n}{\sqrt{2}} \right]$  small squares, call this a  $\mathcal{S}^*$  partition. Observe that all the motes in any small square are at most a distance of  $r_n$  apart. Since there are  $2n / \ln n$  squares and  $n$  motes, by the pigeon hole principle at least one of the small squares

has  $\ln n/2$  motes. Consider this square which is a subgraph  $H$  of  $G_n$ . For each mote  $i$  in  $H$  with sector  $S_i$ , consider the sector  $S_i^* = 2\pi - S_i$ . It has an amplitude of  $\pi - \epsilon$  (see Figure 2.6). That is, the sector which each mote does not see, equals  $\pi - \epsilon$ . The motes of any independent set in  $H$  must form a clique in  $H^*$ , where  $H^*$  is the sector graph induced by  $S_i^*$ . Since the amplitude is less than  $\pi$ , this set must form a convex polygon (as was the case for the clique of  $G_n$  when  $\alpha < \pi - \epsilon$ ), thus  $w(H^*) \leq \lfloor \frac{2\pi}{\alpha - \pi} \rfloor$ . Let  $\vartheta(H)$  represent the independence number of  $H$ . Then  $\vartheta(H) = w(H^*) \leq \lfloor \frac{2\pi}{\alpha - \pi} \rfloor$ . Using the fact that  $\chi(G_n) \geq \chi(H) \geq V_H/\vartheta(H)$ , we have  $\chi(G_n) \geq \frac{\ln n}{2 \lfloor \frac{2\pi}{\alpha - \pi} \rfloor}$ . In order to establish the upper bound we use Brook's Theorem (see Lemma 1.3 in [28]):  $\chi(G_n) \leq \Delta(G_n) + 1$ . Form [24], we know w.h.p.,  $\Delta(G_n)$  is  $\Theta(\ln n)$ . Thus, w.h.p.  $\chi(G_n)$  is  $\Theta(\ln n)$ .  $\square$

## 2.6.2 $\epsilon < \alpha < \pi - \epsilon$

### Proof

- First we establish the lower bound. Note that  $\widehat{\omega}_2(G_n) \leq \chi(G_n)$ , where  $\widehat{\omega}_2(G_n)$  is the size of the maximum undirected clique. Consider the dissection  $\mathcal{S}$  of  $[0, 1]^2$ . Divide each square into  $\ln n$  equal regions by placing  $\ln n$  equally spaced horizontal lines, i.e. a horizontal subdivision of  $\mathcal{S}$  (see Figure 2.3 (b)). By Lemma 3, w.h.p. there is a region  $R_i$  which contains  $(1 - o(1))a_n/B$  motes from each class in  $\mathcal{B}$ . Consider the motes in the region  $R_i$  which belong to class  $B_1$ . Subdivide  $R_i$  into  $a_n/B$  rectangles, by drawing  $a_n/B$  evenly spaced vertical lines. Thus each rectangle has height equal to  $1/b_n$  and width equal to  $\frac{Br_n}{a_n}$  (see Figure 2.7). The expected number of rectangles containing at least one vertex in the limit as  $n \rightarrow \infty$  is  $(1 - \frac{1}{e})a_n/B$ ; again one can show concentration around this expectation via Talagrand's inequality, thus w.h.p. there at least  $(1 - (\frac{1}{e} + o(1)))a_n/B$  such rectangles. Assume (the worst case) a mote  $i$  is in the upper (or lower) right-hand corner of a rectangle, and its orientation angle  $\beta_i = 0$ , after a distance of  $\left\lceil \frac{\cos(\alpha/2)}{\sin(\alpha/2)b_n} \right\rceil$  in the horizontal direction, the mote will be able to see a distance of  $1/b_n$  in the vertical direction. That is after this distance the mote will have an arc with every other mote to

its right within a distance of  $r_n$  in the rectangle in question. Repeating the same arguments as in the case of  $\omega(G_n)$  (which we omit in the interest of space), one can establish, w.h.p.,  $\widehat{\omega}_2(G_n)$  is at least  $d \cdot a_n$ , where  $d$  is some constant dependent on  $\alpha$ .

- Next we establish the upper bound. Consider the dissection  $\mathcal{S}$  on  $[0, 1]^2$ . Let  $\chi^*$  represent the largest chromatic number of any square  $S$  in  $\mathcal{S}$ , then it is easily verifiable  $\chi(G_n)$  is upper-bounded by  $9\chi^*$ , thus in order to upperbound  $\chi(G_n)$ , we upperbound  $\chi^*$ .

Again fix a square  $S$  in  $\mathcal{S}$  and consider the partition  $\mathcal{B}$ . By the Pigeon hole principle, at least one class  $B_j$  will contain  $\frac{d}{B}a_n$  motes in  $S$  all them oriented in almost the same direction. Let  $M_j$  be the set of all such motes. Define a partition of  $S$  into  $\ln n$  strips in the following way. Imagine there exists a mote in  $S$ , whose bisectrix falls exactly in the center of the class  $B_j$ . Draw a line perpendicular to the bisectrix of this (imaginary) mote. This perpendicular line will be the *orientation* of the strips (note in the case of  $\omega(G_n)$ , Theorem 1, a different orientation was used). Partition  $S$  into  $\ln n$  strips parallel to the orientation (see Figure 2.8), in this partition of  $S$ , all the motes in  $M_j$  look in the same *approximate* direction. We wish to prove that for a sufficiently large constant  $d$ , w.h.p. every set of  $\frac{d}{B}a_n$  motes contains an independent set of size at least  $1/3 \ln \ln n$ .

Using similar arguments as in Section 1, one can show w.h.p. at least  $\frac{d}{2B}a_n$  motes fall into a strip having average height  $> \Theta\left(\frac{1}{\sqrt{n}}\right)$ . Thus we only consider motes falling into the strips having average height  $> \Theta\left(\frac{1}{\sqrt{n}}\right)$ . Next, we will order these motes going along the specified direction. For example assume the specified direction is going left to right. Then we label the leftmost mote, 1, the second leftmost mote, 2, and so on. Next we partition the  $\frac{d}{2B}a_n$  motes into  $\frac{d \cdot a_n}{2B \ln \ln n}$  classes. Each class  $C$  will contain  $\ln \ln n$  motes. Again imagine that we are going from left to right, then class one will contain mote<sub>1</sub> to mote <sub>$\ln \ln n$</sub> , and class two will contain the next  $\ln \ln n$  motes. Now again by the pigeon hole principle at least one half of these classes occupy at most  $2 \ln n \frac{2B \ln \ln n}{d \cdot a_n}$  strips. For  $d$  sufficiently large this means at least  $\frac{d \cdot a_n}{4B \ln \ln n}$  classes occupy at most  $2 \ln \ln n$  strips.

Chapter 2. The chromatic number of random scaled sector graphs

Now we consider one class of these motes, say class one. We define two edges to be independent of each other if they have no endpoints in common. Thus the edges  $a-b$  and  $b-c$  are not independent, whereas the edges  $a-b$  and  $c-d$  are independent (where  $a, b, c,$  and  $d$  are vertices).

**Lemma 5** *For any class  $C$  of  $\ln \ln n$  motes, if the largest independent edge set is less than  $1/3 \ln \ln n$ , then there exists an independent set of size  $1/3 \ln \ln n$  or greater.*

**Proof** This follows from the fact that the size of the vertex cover is at most 2 times the size of the maximal independent edge set with minimum cardinality. More specifically, assume the largest set of independent edges in  $C$  is  $\leq 1/3 \ln \ln n$ . Remove all the endpoints (along with any of their edges) from the graph. Since this was the largest independent set of edges (i.e. it is trivially maximal), any other edge not in this set must be dependent relative to some edge in this set (otherwise we would have included it in the set). Thus by removing all the endpoints in this set (the one with the most independent edges) we have deleted all the edges in this subgraph (i.e. in the class in question). Each independent edge has two endpoints. The largest such set is by assumption at most  $1/3 \ln \ln n$ . Thus we have removed at most  $2/3 \ln \ln n$  motes (i.e. vertices). The class to begin with had  $\ln \ln n$  motes. Thus we are left with at least  $1/3 \ln \ln n$  motes. And all the edges have been removed, hence these  $1/3 \ln \ln n$  motes form an independent set and the lemma is proved.  $\square$

Define  $p_3$  to be the probability in any one particular class an independent edge set of size  $1/3 \ln \ln n$  or greater exists. First we will restrict ourselves to classes  $C'$  which occupy  $2 \ln \ln n$  strips or less (at least half of the classes are of this type). Thus every mote is within  $2 \ln \ln n$  strips of any other mote in the class. In order for two motes to share an edge, one mote must see the other going along the specified (in our case going from left to right) direction. There are  $\binom{\ln \ln n}{2} < (\ln \ln n)^2$  possible total edges in the class. Recalling the strips have average height of at least  $\Theta\left(\frac{1}{\sqrt{n}}\right)$ , the probability any one edge is present is  $< \Theta\left(\frac{\ln \ln n}{(\ln n)^{1/2}}\right)$ . The probability any edge exists is independent of any other edge existing (since we are considering independent edge sets). The cardinality of the largest independent edge set is  $\leq 1/2 \ln \ln n$ , thus the total number

of ways to chose an independent set greater than  $1/3 \ln \ln n$  is  $< \ln \ln n^{\ln \ln n}$ . Hence by Boole' inequality for  $d$  sufficiently large,

$$p_3 \leq \Theta \left( \ln \ln n^{\ln \ln n} \left( \frac{\ln \ln n}{(\ln n)^{1/2}} \right)^{1/3 \ln \ln n} \right) \leq \frac{1}{(\ln n)^{4/10}} \stackrel{1/4 \ln \ln n}{=} e^{-\frac{d \ln n}{40B}} .$$

Let  $p_4$  denote the probability in any small square a set with  $d \cdot a_n$  motes does not have an independent set of size  $1/3 \ln \ln n$  or greater. There are at most (w.h.p.)  $n^3$  ways to choose a set of  $d \cdot a_n$  motes in any small square. We are considering  $\frac{d \cdot a_n}{4B \ln \ln n}$  classes of motes, i.e. all the classes which occupy at most  $2 \ln \ln n$  strips. No two classes have any motes in common, thus they are independent of each other. And we have  $n/\ln n$  small squares, so for  $d$  sufficiently large, by Boole's inequality

$$p_4 \leq n^4 e^{-(1/10(\ln \ln n)^2) \left( \frac{d \ln n}{4B(\ln \ln n)^2} \right)} \equiv n^4 e^{-\frac{d \ln n}{40B}} .$$

Therefore, there exists a sufficiently large constant  $d$  such that  $p_4 \rightarrow 0$ .

Thus w.h.p., in every small square any set of  $d \cdot a_n$  motes, has an independent set of size at least  $1/3 \ln \ln n$ . Now take any small square, and keep on choosing independent sets of size  $1/3 \ln \ln n$ . Assign all the motes in the same independent set the same color. When there are less than  $d \cdot a_n$  motes left, assign all the remaining motes a different color. Thus we have colored all the motes in any small square (w.h.p.) with at most  $\frac{3 \ln n - d \cdot a_n}{1/3 \ln \ln n} + d \cdot a_n$  colors. Since the chromatic number of the graph ( $\chi(G_n)$ ), is at most a constant times this amount, w.h.p.  $\chi(G_n) = O\left(\frac{\ln n}{\ln \ln n}\right)$ . Combining with our lower bound, we have  $\chi(G_n)$  is, w.h.p.,  $\Theta\left(\frac{\ln n}{\ln \ln n}\right)$ .  $\square$

## 2.7 Proof of Theorem 3

### 2.7.1 $\alpha > \pi + \epsilon$

**Proof** First we prove for  $\alpha > \pi + \epsilon$ , w.h.p.,  $\widehat{\omega}_2 \geq \Theta(\ln n)$ . Again, consider the dissection  $\mathcal{S}$  of  $[0, 1]^2$ . By the pigeon hole principle at least one of the squares has  $\ln n/2$  motes. Further all

the motes in this square are at most a distance of  $r_n$  apart. Consider the subgraph  $H$  induced by the motes in this square  $S$  and consider the partition  $\mathcal{B}$ . The expected number of motes in any class  $B_i \in \mathcal{B}$  is  $\frac{\ln n}{2B}$ . By Lemma 3, w.h.p. every class contains  $(1 - o(1))\frac{\ln n}{2B}$  motes. Next Divide  $S$  into  $\ln n/2B$  stripes, by drawing  $\ln n/2B$  evenly spaced vertical lines. The expected number of strips containing at least one vertex in the limit as  $n \rightarrow \infty$  is  $(1 - \frac{1}{e})\frac{\ln n}{2B}$ ; and by Talagrand's inequality w.h.p., there at least  $(1 - (\frac{1}{e} + o(1)))\frac{\ln n}{2B}$  such strips. Consider the motes in  $B_1$ , going from left to right, every mote can see every other mote to its right (since  $\alpha > \pi$ ). Thus, w.h.p.,  $\widehat{\omega}_2$  is at least  $(1 - (\frac{1}{e} + o(1)))\frac{\ln n}{2B}$ .

For the upper bound, we know w.h.p.,  $\Delta(G_n) < \Theta(\ln n)$ . □

### 2.7.2 $\alpha < \pi - \epsilon$

**Proof** We already established in our proof of  $\chi(G_n)$  (see 5.2) a lower bound of  $\Theta(\frac{\ln n}{\ln \ln n})$ . For the upper bound, the proof is similar to that for  $\omega(G_n)$  (see 4.2), and it is omitted, thus w.h.p.,  $\widehat{\omega}_2 \leq \Theta(\frac{\ln n}{\ln \ln n})$ . □

## 2.8 Conclusions and open problems

In this work, we determined asymptotic values for the directed clique  $\omega^*(G_n)$ , the modified clique  $\widehat{\omega}_2(G_n)$  and the chromatic number  $\chi(G_n)$  of random scaled sector graphs. We observed  $\omega^*$  exhibits a threshold at  $\alpha = 2\pi$  and at  $\alpha = \pi$ , but we have been unable to compute the value of  $\omega^*$  for the particular value of  $\alpha = \pi$ . Similarly, there are thresholds in the behavior of  $\widehat{\omega}_2(G_n)$  and  $\chi(G_n)$  at  $\alpha = \pi$ . Again, our methods do not seem to work for computing  $\widehat{\omega}_2(G_n)$  and  $\chi(G_n)$ , for the particular value of  $\alpha = \pi$  and the computation of  $\widehat{\omega}_2(G_n)$ ,  $\chi(G_n)$  and  $\omega^*(G_n)$  at  $\alpha = \pi$  remain open problems.

In the framework of channel assignment on random geometric graphs, McDiarmid [27]

## Chapter 2. The chromatic number of random scaled sector graphs

studied the value of the chromatic number for two types of random geometric graphs. In the case of sparse graphs, defined as  $r_n^2 n / \ln n \rightarrow 0$ , (i.e. the average degree is  $o(\ln n)$ ), in probability  $\chi/\omega \rightarrow 1$ . In the case of dense graphs, defined as  $r_n^2 n / \ln n \rightarrow \infty$ , in probability  $\chi/\omega \rightarrow 2\sqrt{3}/\pi \sim 1.103$ . McDiarmid leaves as open the problem of computing the chromatic number of random geometric graphs for which  $r_n^2 n / \ln n$  tends to a constant  $c$  as  $n \rightarrow \infty$ . Notice this radii is the one considered in this chapter for random scaled sector graphs. Thus the work of McDiarmid suggests two lines for further investigation; the asymptotic evaluation of  $\chi(G_n)/\widehat{\omega}_2(G_n)$  when  $r_n^2 n / \ln n$  tends to a constant  $c$ , and the study of  $\chi(G_n)$  and  $\widehat{\omega}_2(G_n)$  for the particular cases of *sparse* and *dense* random scaled sector graphs.

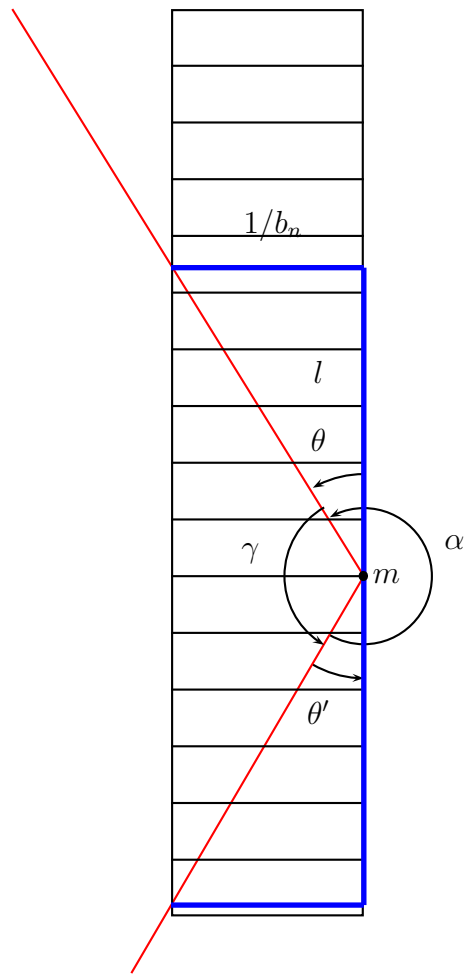


Figure 2.4: Proof of lower bound

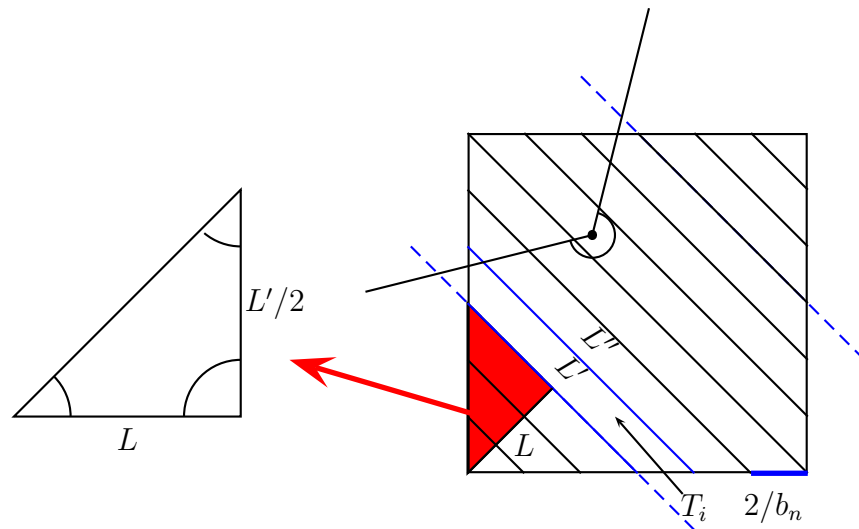


Figure 2.5: Partition of  $S$  by strips

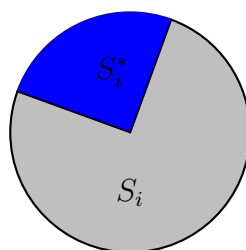


Figure 2.6: Sector  $S_i$  and complementary sector  $S_i^*$

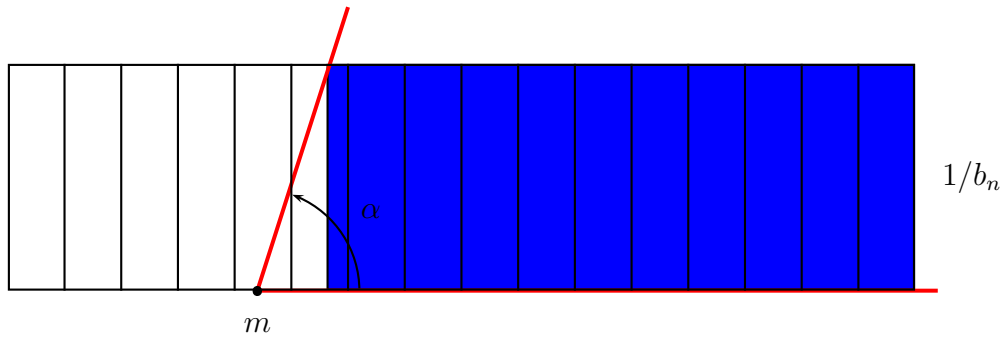


Figure 2.7: Figure for the proof of 5.2

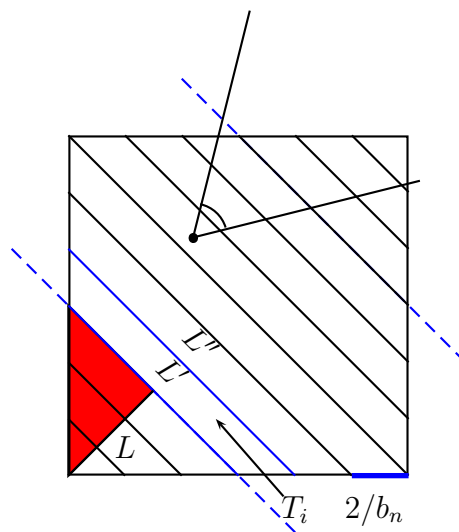


Figure 2.8: Partition of  $S$  in the prove 5.2

## Chapter 3

# Counting Connected Graphs and Hypergraphs via the Probabilistic Method

### 3.1 Acknowledgements

I would like to thank Amin Coja-Oghlan and Cristopher Moore. The work in this chapter was done jointly with them, a preliminary version of which appeared in [35].

### 3.2 Introduction and Results

A  $d$ -uniform hypergraph  $H$  consists of a set  $V$  of vertices and a set  $E$  of edges, which are subsets of  $V$  of cardinality  $d$ . For instance, a 2-uniform hypergraph is just a graph. A vertex  $w$  is *reachable in  $H$*  from a vertex  $v$  if either  $v = w$  or there is a sequence  $e_1, \dots, e_k$  of edges such that  $v \in e_1$ ,  $w \in e_k$ , and  $e_i \cap e_{i+1} \neq \emptyset$  for  $i = 1, \dots, k - 1$ . Of course, reachability in  $H$  is an equivalence relation. The equivalence classes are the *components* of  $H$ , and  $H$  is *connected* if there is only one component.

Connectedness is perhaps the most basic property of graphs and hypergraphs, and therefore estimating the number of connected graphs or hypergraphs with a given number of vertices and edges is a fundamental combinatorial problem. In most applications, one is interested in asymptotic results, where the number of vertices/edges tends to infinity. The main result in this chapter is a formula for the asymptotic number of connected  $d$ -uniform hypergraphs; or, equivalently, an estimate of the probability that a random  $d$ -uniform hypergraph is connected. Our results hold up to a constant multiplicative factor. We study both a model  $H_d(n, m)$  in which the number of vertices and edges is fixed, and a binomial model  $H_d(n, p)$  in which each possible edge appears with probability  $p$  independently. In the latter case we also calculate the expected number of edges given that the hypergraph is connected. Furthermore, we obtain a simple algorithm for generating a connected hypergraph uniformly at random.

For the special case of graphs, i.e., for  $d = 2$ , the results we present in this chapter (and, in some cases, stronger results) are already known. By contrast, little has previously been known about  $d$ -uniform hypergraphs where  $d \geq 3$ , although a few papers deal with the component structure of random hypergraphs. We will discuss related work in Section 3.2.3 after presenting our main results.

We obtain our results using a new probabilistic approach. Rather than using powerful techniques of enumerative combinatorics such as generating functions, our calculations are for the most part elementary, and just rely on the fact that each connected (hyper)graph of a given order and size is equally likely to occur as the “giant component” of a larger one. We believe that this approach is of interest in its own right, and that similar ideas can be applied to a number of further problems in combinatorics and random (hyper)graph theory.

### 3.2.1 Results

Throughout, we will consider  $d$ -uniform hypergraphs where  $d \geq 2$  is a fixed integer. Given a hypergraph  $H = (V, E)$ , its *order* is its number of vertices  $|V|$ , and its *size* is its number of edges

$|E|$ . The *degree* of a vertex  $v \in V$  is the number of edges  $e \in E$  such that  $v \in e$ . Hence, if  $H$  is  $d$ -uniform, its average degree is  $d|E|/n$ . Observe that if a  $d$ -uniform hypergraph is connected, then its average degree is at least  $(1 - n^{-1})\frac{d}{d-1}$ .

Two natural random models of hypergraphs present themselves. In  $H_d(n, m)$ , we select one of the  $\binom{n}{m}$  sets of  $m$  possible edges uniformly at random. In  $H_d(n, p)$ , each of the  $\binom{n}{d}$  edges appears independently with probability  $0 \leq p \leq 1$ , in which case the expected number of edges is  $\binom{n}{d}p$  and the expected average degree is  $\binom{n-1}{d-1}p$ . Thus we will compare  $H_d(n, m)$  where  $m = cn/d$  with  $H_d(n, p)$  where  $p = c/\binom{n-1}{d-1}$ . While these two models are interchangeable in many respects, we will show that their probabilities of connectedness differ by an exponential factor if the average degree is constant. We say that  $H_d(n, m)$  or  $H_d(n, p)$  has a certain property *with high probability* (w.h.p.) if the probability that the property holds tends to 1 as  $n \rightarrow \infty$ .

The following theorem determines the asymptotic probability of connectedness of  $H_d(n, m)$ , or equivalently the number of connected  $d$ -uniform hypergraphs, up to a constant factor in the regime where the average degree is  $d/(d-1) + \Omega(1) < c = o(\ln n)$ . We let  $c_d(n, m)$  signify the probability that the random hypergraph  $H_d(n, m)$  is connected, and in addition  $C_d(n, m)$  denotes the number of connected  $d$ -uniform hypergraphs of order  $n$  and size  $m$ .

**Theorem 1** *Let  $c_0 > d/(d-1)$  be a number independent of  $n$ . Then there exist numbers  $C_0, C'_0$  that depend only on  $c_0$  such that for all  $c_0 \leq c = c(n) = o(\ln n)$  the following holds. The equation*

$$1 - a = \exp \left( -ca \cdot \frac{1 - (1 - a)^{d-1}}{1 - (1 - a)^d} \right) \quad (3.1)$$

*has a unique solution  $a = a(c)$  in the interval  $(0, 1)$ . Set  $m = cn/d$ . Then for all sufficiently large  $n$*

$$\begin{aligned} C_0 \Phi_d(c)^n &\leq c_d(n, m) = C_d(n, m) \binom{\binom{n}{d}}{m}^{-1} \leq C'_0 \Phi_d(c)^n, \text{ where} \\ \Phi_d(c) &= a^{1-c} (1 - a)^{(1/a)-1} (1 - (1 - a)^d)^{c/d}. \end{aligned}$$

The next theorem gives an analogous result for  $H_d(n, p)$ . We let  $c_d(n, p)$  signify the probability that  $H_d(n, p)$  is connected.

**Theorem 2** *For each constant  $c_0 > 0$  there exists numbers  $C_1, C'_1$  depending only on  $c_0$  such that for all  $c_0 \leq c = c(n) = o(\ln n)$  the following holds. Let  $p = c/\binom{n-1}{d-1}$ . Then the equation*

$$1 - a = \exp\left(-c \cdot \frac{1 - (1 - a)^{d-1}}{a^{d-1}}\right) \quad (3.2)$$

*has a unique solution  $a$  in the interval  $(0, 1)$ . For all sufficiently large  $n$  we have*

$$C_1 \Psi_d(c)^n \leq c_d(n, p) \leq C'_1 c_d(n, p), \text{ where } \Psi_d(c) = a(1-a)^{(1/a)-1} \exp\left(\frac{c}{d} \cdot \frac{1 - a^d - (1-a)^d}{a^d}\right).$$

*For  $d = 2$  the formula simplifies to  $\Psi_2(c) = 1 - \exp(-c)$ .*

Note  $H_d(n, p)$  is exponentially more likely to be connected than  $H_d(n, m)$  if  $m = \binom{n}{d}p$ . The reason for this is that in  $H_d(n, p)$  the number of edges is a random variable: we can think of  $H_d(n, p)$  as first choosing a number of edges  $m'$  according to the binomial distribution  $\text{Bin}(\binom{n}{d}, p)$ , and then choosing a  $H_d(n, m')$ . Thus  $H_d(n, p)$  can boost its probability of being connected by including a larger number  $m' > \binom{n}{d}p$  of edges. Indeed, given that  $H_d(n, p)$  is connected, the conditional expectation of the number of edges will be significantly larger than  $\binom{n}{d}p$ . Our next theorem quantifies this observation.

**Theorem 3** *Let  $c, a,$  and  $p$  be as in Theorem 2. Set  $\gamma_d(c) = c [1 - (1 - a)^d] a^{-d}$ . Then the expected number of edges of  $H_d(n, p)$  given that that  $H_d(n, p)$  is connected is  $n\gamma_d(c)/d + o(n)$ , so the expected average degree is  $\gamma_d(c) + o(1)$ . For  $d=2$ , the formula simplifies to  $\gamma_2(c) = c \cdot \coth(c/2)$ .*

As  $c \rightarrow \infty$ ,  $\gamma_d(c) \sim c$ , since connectedness becomes a less unusual condition as  $c$  increases. As  $c \rightarrow 0$ ,  $\gamma_d(c) \rightarrow d/(d - 1)$ , the minimum average degree required for connectivity.

Theorem 1 addresses the combinatorial problem of estimating the number of connected hypergraphs. But there is also a corresponding algorithmic problem: can we sample a connected

hypergraph of a given order and size efficiently uniformly at random? We answer this question in the affirmative.

**Theorem 4** *Let  $c_0 > d/(d-1)$  be fixed, let  $c_0 \leq c = o(\ln n)$ , and let  $m = cn/d$ . There is a randomized algorithm that samples a connected hypergraph of order  $n$  and size  $m$  uniformly at random in expected time  $O(n^{1/2}m^{3/2})$ .*

One ingredient of the proofs of Theorems 1–4 is a result on the component structure of random hypergraphs  $H_d(n, m)$  and  $H_d(n, p)$ . Suppose that the average degree  $c = dm/n$  (resp.  $c = \binom{n-1}{d-1}p$ ) is a constant greater than  $1/(d-1)$ . Then, just as for graphs (cf. [34, 39]), w.h.p. there is a unique *giant component* of order  $\Omega(n)$  in  $H_d(n, p)$ , while all other components are of order  $O(\ln n)$ . To prove Theorems 1–4, it is crucial to have a rather tight estimate on the order and size of the giant component. Since these considerations may be of independent interest, we state the result as a theorem. For a hypergraph  $H$  we let  $\mathcal{N}(H)$  (resp.  $\mathcal{M}(H)$ ) denote the largest order (size) of a component of  $H$ .

**Theorem 5** *Let  $p = c \binom{n-1}{d-1}^{-1}$  and  $m = \binom{n}{d}p = cn/d$ .*

1. *If there is a constant  $c_0 < (d-1)^{-1}$  such that  $c \leq c_0$ , then w.h.p. we have  $\mathcal{N}(H_d(n, p)) = O((1 - (d-1)c_0)^{-2} \ln n)$ .*
2. *Suppose that  $c_0 > (d-1)^{-1}$  is a constant, and that  $c_0 \leq c = o(\ln n)$ . There is a unique number  $0 < a = a(c) < 1$  such that  $1 - a = \exp[c((1-a)^{d-1} - 1)]$ . Set  $b = b(c) = 1 - (1-a)^d$ . Then we have*

$$1 - a = \exp(-\Theta(c)), \quad (3.3)$$

$$\mathbb{E}[\mathcal{N}(H_d(n, p))] = an + n^{o(1)}, \quad \mathbb{E}[\mathcal{M}(H_d(n, p))] = bm + n^{o(1)}, \quad (3.4)$$

$$\text{Var}(\mathcal{N}(H_d(n, p))) \sim \frac{a(1-a) \left[ 1 + ((d-1)c - 1) \sum_{i=1}^{d-2} (1-a)^i \right]}{(1 - c(d-1))(1-a)^{d-1}} \cdot n. \quad (3.5)$$

Furthermore, with probability  $\geq 1 - n^{-10}$  there is precisely one component of order  $(1 + o(1))an$  in  $H_d(n, m)$ , while all other components have order  $O(c((d-1)c-1)^{-2} \ln n)$ . In addition, with probability  $1 - \exp(-n^{\Omega(1)})$  the number of isolated vertices is  $n \exp(-c) + o(n^{9/10})$ . Finally, for each  $\varepsilon > 0$  there is a constant  $C_\varepsilon > 0$  such that

$$\mathbb{P} [|\mathcal{M}(H_d(n, p)) - bm| \leq C_\varepsilon n^{d/2} p^{1/2}] \geq 1 - \varepsilon. \quad (3.6)$$

### 3.2.2 Techniques and Overview

While most of the previous work on counting connected (hyper)graphs relies on rather heavy machinery from enumerative combinatorics, our approach is quite different and employs only relatively simple probabilistic techniques. For instance, the basic idea behind the proof of Theorem 1 is as follows. Suppose that  $p = c/\binom{n-1}{d-1}$  for some  $c > 1/(d-1)$ , and let  $a, b$  be as in Theorem 5. Then, by Theorem 5, w.h.p. there is a unique giant component in  $H_d(n, p)$ , which—conditioned on its order and size—is a uniformly random connected hypergraph. To prove Theorem 1, we derive an explicit formula that expresses the number  $C_d(\nu, \mu)$  of connected hypergraphs of order  $\nu = an$  and size  $\mu = bm$  in terms of the probability  $\chi(\nu, \mu)$  that the giant component of  $H_d(n, p)$  has precisely order  $\nu$  and size  $\mu$ . Then, we prove that  $\chi(\nu, \mu) = \Theta((1-a)bm)^{-1/2}$  to obtain the estimate of  $C_d(\nu, \mu)$  stated in Theorem 1.

To compute  $\chi(\nu, \mu)$ , we expose the edges of  $H_d(n, p)$  in two rounds. In the first round we include edges with probability  $p_1 = (1-\varepsilon)p$  independently, where  $\varepsilon > 0$  is a sufficiently small constant. In the second round we add further edges with probability  $p_2$ , where  $p_1 + p_2 - p_1 p_2 = p$ , so that the resulting hypergraph is distributed as  $H_d(n, p)$ . As Theorem 5 yields the approximate order and size of the giant component of the random hypergraph  $H_1 = H_d(n, p_1)$  generated in the first round, to compute  $\chi(an, bm)$  we just need to study the growth of the giant of  $H_1$  when further edges are added with probability  $p_2$ , i.e, with a fairly small probability.

The (somewhat technical) proof of Theorem 5 is based on the analogy between exploring the components of a random hypergraph  $H_d(n, p)$  and a Galton-Watson branching process. This

analogy is well known for graphs, cf. [30, Chapter 10] and [39, Chapter 5]. While the basic approach we use is the same as in [30, 39], we need to modify the previous techniques in three respects.

- We generalize the branching processes considered in [30, 39] so that they describe exploring the components of  $d$ -uniform hypergraphs rather than graphs.
- In [30, 39] it is shown via the branching process that the expected order and size of the giant component of  $H_2(n, p)$  is  $(1 + o(1))an$  (resp.  $(1 + o(1))bm$ ). By comparison, (3.4) gives the order/size of the giant up to an additive error of  $n^{o(1)}$ . To obtain these more precise statements, more involved technical estimates are necessary.
- Additionally, we also use the branching process to determine the variance of the number of vertices in the giant component and the variance of the number of edges outside the giant component.

In Theorems 1–5 we assume that  $c = o(\ln n)$ . Although our arguments also apply if  $c = \Omega(\ln n)$ , the assumption  $c = o(\ln n)$  is convenient in order to avoid technical case distinctions. Moreover, if  $c > (1 + \varepsilon) \ln n$  for an arbitrarily small but constant  $\varepsilon > 0$ , then w.h.p. both  $H_d(n, p)$  and  $H_d(n, m)$  are connected, i.e.,  $c_d(n, m) = c_d(n, p) = 1$  (the proof given in for the case  $d = 2$  in [34, Section 7.1] carries over to larger  $d$  without essential modifications).

The chapter is organized as follows. In Section 3.3 we introduce some notation and make a few preliminary remarks. Then, in Section 3.4 we derive the formula for  $C_d(\nu, \mu)$  in terms of  $\chi(\nu, \mu)$ . In addition, we also obtain a similar formula for  $c_d(\nu, p)$  to prove Theorem 2, and we prove Theorem 4. Furthermore, in Section 3.5, we determine  $\chi(\nu, \mu)$ . Thus, Sections 3.4–3.5 present the main ideas of this chapter. Then, in Section 3.6 we show how Theorem 1 implies Theorem 3. Finally, Section 3.7 deals with the proof of Theorem 5.

### 3.2.3 Related Work

The asymptotic number of connected *graphs* with a given number of edges and vertices was determined by Bender, Canfield, and McKay [32]. Their proof is based on generating functions, and gives a stronger result than Theorem 1 in that it determines the number of connected graphs up to a multiplicative factor  $1 + o(1)$  instead of a constant. The same authors determined the probability that a random graph  $G(n, p) = H_2(n, p)$  is connected as well as the expected number of edges of  $G(n, p)$  conditioned on connectedness [33], corresponding to the case  $d = 2$  of Theorems 2 and 3. Indeed, they even give the asymptotic distribution of the number of edges conditioned on connectedness, not just its expectation. An algorithm for generating random connected graphs (the case  $d = 2$  of Theorem 4) immediately follows from the results in [32, 33]. Pittel and Wormald [45] presented an alternative approach based on enumerating graphs of minimum degree 2 to derive improved versions of some of the results in [32, 33].

The basic idea of this chapter—expressing  $C_d(\nu, \mu)$  in terms of  $\chi(\nu, \mu)$  and computing the latter via probabilistic methods, cf. Section 3.2.2—is related to the work of Łuczak [42] on bounding  $C_2(\nu, \mu)$  with  $\mu = (1 + o(1))\nu$ . To bound  $C_2(\nu, \mu)$  from above, Łuczak expresses the expected number of components of  $G(n, p) = H_2(n, p)$  of order  $\nu$  and size  $\mu$  in terms of  $C_2(\nu, \mu)$ . Then, he points out that the expected number of such components is trivially bounded by  $n/\nu$ , thereby obtaining an upper bound on  $C_2(\nu, \mu)$ . The difference between Łuczak’s approach and ours is that we actually determine the probability that  $H_d(n, p)$  has a component of order  $\nu = an$  and size  $\mu = bm$  up to a constant factor, rather than using the trivial upper bound  $n/\nu$  on the number of such components. Moreover, we use the probabilistic argument to get both an upper and a lower bound on  $C_d(\nu, \mu)$ , while the lower bound given in [42] relies on different techniques (e.g., on an explicit formula for the asymptotic number of 3-regular graphs).

O’Connell [43] showed that the probability  $c_2(\nu, p)$  that  $G(\nu, p)$  is connected is  $\Psi_2(\nu p)^{\nu+o(\nu)}$  (cf. the  $d = 2$  case of Theorem 2). The proof relies on the fact that w.h.p. the giant component of  $G(n, p)$  has order  $an + o(n)$  where  $1 - a = \exp(-ac)$ ,  $0 < a < 1$  (cf. Theorem 5). However,

since O’Connell avoids computing the probability that the order of the giant component actually equals  $\nu = an$ , he only obtains the value of  $c_2(\nu, p)$  up to a factor of  $\exp(o(\nu))$  (rather than a constant factor as in Theorem 2).

Independently of us and simultaneously, van der Hofstad and Spencer [37] determined the number  $C_2(n, m)$  of connected graphs via probabilistic methods up to a factor  $1 + o(1)$ . The basic ingredient to their analysis is a novel point of view of the branching process argument for exploring the components of  $G(n, p)$ . However, there seems to be no direct way to extend the approach of [37] to hypergraphs.

The asymptotic order and size of the giant component of random graphs (the case  $d = 2$  of Theorem 5) has been known since the pioneering work of Erdős and Renyi, cf. [34, 39] for unified treatments. The distribution of the giant component was further investigated by several researchers, cf. [31, 38, 44] and the monographs [34, 39].

In comparison to the case of graphs ( $d = 2$ ), little is known for  $d$ -uniform hypergraphs with  $d \geq 3$ . Karoński and Łuczak [41] determined the number of connected hypergraphs of order  $n$  and size  $m = n/(d - 1) + k$  where  $k = o(\ln n / \ln \ln n)$  (i.e., just above the number of edges necessary for connectedness) up to a factor of  $1 + o(1)$  via purely combinatorial techniques. Since Theorem 1 addresses the case  $m = cn/d$  for  $c \geq d/(d - 1) + \Omega(1)$ , our results and those of [41] are incomparable.

Moreover, Schmidt-Pruzan and Shamir [46] showed that in a very general model of random hypergraphs a *phase transition* phenomenon occurs: there is a certain average degree  $c^*$  such that for all  $c < (1 - \varepsilon)c^*$  the largest component of  $H_d(n, p)$  with  $\binom{n-1}{d-1}p = c$  has order  $O(\ln n)$ , whereas for  $c > (1 + \varepsilon)c^*$  there is a unique giant component of order  $\Omega(n)$  w.h.p.; here  $\varepsilon > 0$  denotes an arbitrarily small constant. In the case of  $d$ -uniform hypergraphs, the critical average degree is  $c^* = 1/(d - 1)$ . Karoński and Łuczak [40] studied this phase transition in greater detail, considering also the case that  $c \sim c^*$ . To the best of our knowledge, the expected order and size of the giant component as given in Theorem 5 have not been stated explicitly before.

### 3.3 Preliminaries

**Notation.**

For a real number  $x$  and an integer  $r \geq 1$ , we let  $(x)_r = \prod_{j=0}^{r-1} x - j$  denote the  $r$ 'th falling factorial.

Throughout, we let  $V = \{1, \dots, n\}$  be a set of  $n$  labeled vertices. If  $H$  is a hypergraph, then  $V(H)$  denotes the vertex set and  $E(H)$  the edge set of  $H$ .

We investigate  $d$ -uniform hypergraphs, where the number  $d \geq 2$  is constant, i.e., independent of  $n$ . Further, we let  $p = c/\binom{n-1}{d-1}$ , so that the average degree of a random hypergraph  $H_d(n, p)$  is  $c$ . We always assume that  $c = o(\ln n)$ . Unless otherwise specified,  $a = a(c)$  and  $b = b(c)$  signify the numbers defined in Theorem 5 (if  $c > (d-1)^{-1}$ ).

**Asymptotics.**

As we are interested in asymptotic results that hold as  $n \rightarrow \infty$ , in our computations we may always assume that  $n$  is large enough. If  $f(c, n), g(c, n)$  are functions, then we use the common notation  $f = O(g)$ ,  $f = \Omega(g)$ , and  $f = \Theta(g)$  in the sense that the constants hidden in the  $O$ ,  $\Omega$ , or  $\Theta$  are independent of  $c$ . For example,  $f = O(g)$  means that there is a constant  $C$  such that for all  $c$  and all sufficiently large  $n$  we have  $|f(c, n)| \leq Cg(c, n)$ . We use the symbol  $f = \tilde{O}(g)$  if there is a constant  $K > 0$  such that  $|f(c, n)| \leq g(c, n) \ln^K n$  for all  $c$  and all sufficiently large  $n$ .

In addition, the symbols  $O_c$ ,  $\Omega_c$ , and  $\Theta_c$  to indicate asymptotics as  $n \rightarrow \infty$  so that the constants hidden in the  $O_c$ ,  $\Omega_c$ , or  $\Theta_c$  symbol depend on  $|c - (d-1)^{-1}|$ . For instance, we write  $f = \Theta_c(g)$  if for each  $\varepsilon > 0$  there are constants  $C(\varepsilon), C'(\varepsilon)$  such that for all  $c$  satisfying  $|c - c_0| > \varepsilon$  and all sufficiently large  $n$  we have  $C'(\varepsilon)g(c, n) \leq f(c, n) \leq C(\varepsilon)g(c, n)$ . The symbols  $f = O_c(g)$  and  $f = \Omega_c(g)$  are defined similarly.

We shall frequently encounter terms of the type  $(1+x)^y$ , where  $x = x(n)$  and  $y = y(n)$  are

functions of  $n$ . To estimate such these terms, the following two elementary estimates are useful:

$$1 + x = \exp(x + x^2/2 + O(x^3)) = \exp(x + O(x^2)) \quad \text{as } x \rightarrow 0. \quad (3.7)$$

Hence, if  $x^2y \rightarrow 0$  as  $n \rightarrow \infty$ , then (3.7) yields  $(1 + x)^y \sim \exp(xy)$ . Moreover, if  $x^3y \rightarrow 0$ , then  $(1 + x)^y \sim \exp(xy + x^2y/2)$ .

### Probability distributions.

By  $\text{Bin}(\nu, q)$  we denote a binomially distributed random variable with parameters  $\nu$  and  $q$ . Moreover,  $\text{Po}(c)$  signifies a Poisson random variable with mean  $c$ . We will apply the following Chernoff bound on the tails of a binomial or Poisson random variable  $X$  with mean  $\mu$  several times (cf. [39, p. 27] for a proof):

$$\mathbb{P}[X \leq \mu - t] \leq \exp\left[-\frac{t^2}{2\mu}\right], \quad \mathbb{P}[X \geq \mu + t] \leq \exp\left[-\frac{t^2}{2(\mu + t/3)}\right] \quad (t > 0). \quad (3.8)$$

Moreover, the following estimate on the binomial distribution is an important ingredient to the proofs of Theorems 1 and 2.

**Lemma 1** *For any constant  $C > 0$  there exist numbers  $c_0, c_1, c_2$  such that the following holds. Let  $X = \text{Bin}(\nu, q)$ ,  $\mu = \mathbb{E}(X) = \nu q$ , and  $\sigma^2 = \text{Var}(X) = \nu q(1 - q)$ , and suppose that  $\sigma > c_0$ . Then for all integers  $t$  such that  $\mu - C\sigma \leq t \leq \mu + C\sigma$  we have  $c_1\sigma^{-1} \leq \mathbb{P}[X = t] \leq c_2\sigma^{-1}$ .*

**Proof** This follows from a direct evaluation of the term  $\mathbb{P}[X = t] = \binom{\nu}{t} p^t (1 - p)^{\nu - t}$  using Stirling's formula.  $\square$

Loosely speaking, Lemma 1 states that for any binomially distributed  $X$  with sufficiently large standard deviation  $\sigma$  each value that deviates from  $\mu = \mathbb{E}(X)$  by at most  $C\sigma$  occurs with ‘‘approximately’’ uniform probability  $\Theta(\sigma^{-1})$ . Let us also recall the following simple estimate for  $X = \text{Bin}(\nu, q)$ , which easily follows from Stirling's formula:

$$\forall 0 \leq t \leq \nu : \mathbb{P}[X = t] = O(\sigma^{-1}), \quad \text{where } \sigma^2 = \nu q(1 - q). \quad (3.9)$$

Finally, we let  $\text{Bin}_{\geq 1}(\nu, q)$  be a binomially distributed random variable  $X$  with parameters  $\nu$  and  $q$ , given that  $X \geq 1$ . That is,  $\mathbb{P}[X = j] = [1 - (1 - q)^\nu]^{-1} \binom{\nu}{j} q^j (1 - q)^{\nu-j}$  for  $1 \leq j \leq \nu$ . The mean and the variance of  $X$  are easily computed:

$$\mathbb{E}[X] = \frac{\nu q}{1 - (1 - q)^\nu}, \quad \text{Var}[X] = \frac{nq(1 - q)}{1 - (1 - q)^\nu} + \frac{(1 - q)^\nu n^2 q^2}{(1 - (1 - q)^\nu)^2}. \quad (3.10)$$

Recall that a random variable  $X$  *dominates* another random variable  $Y$  if for all real  $t$  we have  $\mathbb{P}[X \leq t] \leq \mathbb{P}[Y \leq t]$ . In particular, if  $X$  dominates  $Y$  then  $\mathbb{E}(X) \geq \mathbb{E}(Y)$ .

## 3.4 The Number of Connected Hypergraphs

### 3.4.1 Outline

Throughout, we assume that  $c_0 \leq c = o(\ln n)$  for some constant  $c_0 > (d - 1)^{-1}$ . We set  $p = c/\binom{n-1}{d-1}$ ,  $m = cn/d$ , and let  $a, b$  be as in Theorem 5. In this section we prove Theorems 1–2 and 4, assuming Theorem 5. More precisely, we reduce the problem of estimating the probability that  $H_d(n, m)$  or  $H_d(n, p)$  is connected to the problem of computing the probability that the giant component of  $H_d(n, p)$  has *exactly* order  $an$  and size  $bm$ , where  $a, b$  are as in Theorem 5. Let us first outline the proofs, deferring most of the technical details to Sections 3.4.2–3.4.4.

#### Proof of Theorem 1.

Let  $\chi(an, bm) = \mathbb{P}[\mathcal{N}(H_d(n, p)) = an \wedge \mathcal{M}(H_d(n, p)) = bm]$  be the probability that in a random hypergraph  $H_d(n, p)$  the order and the size of the giant component equal *exactly*  $an$  and  $bm$ . In order to prove Theorem 1, we shall establish that

$$\chi(an, bm) \sim \binom{n}{an} C_d(an, bm) p^{bm} (1 - p)^{\binom{n}{d} - \binom{1-a}{d}n - bm}. \quad (3.11)$$

Then, we will compute  $\chi(an, bm)$ , so that finally we can just solve (3.11) for  $C_d(an, bm)$  to obtain the formula stated in Theorem 1.

Chapter 3. Counting Connected Graphs and Hypergraphs via the Probabilistic Method

To bound  $\chi(an, bm)$  from above, let us first consider the *expected* number of components of order  $an$  and  $bm$  occurring in  $H_d(n, p)$ . There are  $\binom{n}{an}$  ways to choose a set  $S$  of  $an$  vertices where to place a component of order  $an$ . Then, there are  $C_d(an, bm)$  ways to choose a connected hypergraph  $\mathcal{H}$  of order  $an$  and size  $bm$ , and the probability that the subhypergraph of  $H_d(n, p)$  induced on  $S$  is  $\mathcal{H}$  equals  $p^{bm}(1-p)^{\binom{an}{d}-bm}$ , because each of the  $\binom{an}{d}$  possible edges inside of  $S$  is present with probability  $p$  independently. Finally, if  $S$  is a component of  $H_d(n, p)$ , then none of the  $\binom{n}{d} - \binom{an}{d} - \binom{(1-a)n}{d}$  possible  $S-V \setminus S$  edges is present in  $H_d(n, p)$ . As each of these edges occurs with probability  $p$  independently, the expected number of components of order  $an$  and size  $bm$  is

$$\binom{n}{an} C_d(an, bm) p^{bm} (1-p)^{\binom{n}{d} - \binom{(1-a)n}{d} - bm} \geq \chi(an, bm). \quad (3.12)$$

Furthermore,  $\chi(an, bm)$  is bounded from below by the probability that  $H_d(n, p)$  contains *precisely one* component of order  $an$  and size  $bm$ . Hence, adding to (3.12) a further factor to ensure that  $H_d(n, p) - S$  has no large component, we get

$$\binom{n}{an} C_d(an, bm) p^{bm} (1-p)^{\binom{n}{d} - \binom{(1-a)n}{d} - bm} \cdot \mathbb{P}[\mathcal{N}(H_d((1-a)n, p)) = O_c(\ln n)] \leq \chi(an, bm). \quad (3.13)$$

In Section 3.4.2 we shall prove that the additional factor is indeed negligible.

**Lemma 2** *We have*  $\mathbb{P}[\mathcal{N}(H_d((1-a)n, p)) = O_c(\ln n)] \sim 1$ .

Combining (3.12), (3.13), and Lemma 2, we obtain (3.11).

In the light of (3.11), the crucial step in the proof of Theorem 1 is to estimate  $\chi(an, bm)$ . We shall carry out this estimate in Section 3.5, proving the following proposition.

**Proposition 1** *We have*  $\chi(an, bm) = \Theta_c((1-a)bnm)^{-1/2}$ .

Combining (3.12) and Proposition 1 and invoking Stirling's formula, we obtain

$$\begin{aligned}
 \Theta_c [(1-a)bnm]^{-1/2} &= \binom{n}{an} \binom{\binom{an}{d}}{bm} c_d(an, bm) p^{bm} (1-p)^{\binom{n}{d} - \binom{(1-a)n}{d} - bm} \\
 &= c_d(an, bm) \cdot \Theta [a(1-a)bnm]^{-1/2} a^{-an} (1-a)^{-(1-a)n} \left[ \binom{an}{d} p (bm)^{-1} \right]^{bm} \\
 &\quad \times \left[ \binom{an}{d} \left[ \binom{an}{d} - bm \right]^{-1} \right]^{\binom{an}{d} - bm} (1-p)^{\binom{n}{d} - \binom{(1-a)n}{d} - bm}. \quad (3.14)
 \end{aligned}$$

To estimate the terms occurring in (3.14), some tedious computations are needed, which we will carry out in Section 3.4.3. The following lemma summarizes the results.

**Lemma 3** 1. We have  $\left[ \binom{an}{d} p (bm)^{-1} \right]^{bm} = \Theta(1) \cdot \left( \frac{a^d}{b} \right)^{bm}$ .

2. Moreover,  $\left[ \binom{an}{d} \left[ \binom{an}{d} - bm \right]^{-1} \right]^{\binom{an}{d} - bm} (1-p)^{\binom{n}{d} - \binom{(1-a)n}{d} - bm} = \Theta_c(1)$ .

Plugging the estimates from Lemma 3 into (3.14) and solving for  $c_d(an, bm)$ , we get

$$c_d(an, bm) = \Theta(1) \cdot [a(1-a)^{(1-a)/a}]^{an} [a^{-d}b]^{bm}. \quad (3.15)$$

Further, setting  $\nu = an$ ,  $\mu = bm$ , and  $\zeta = d\mu/\nu = bc/a$  and recalling that  $b = 1 - (1-a)^d$ , we can rewrite (3.15) as

$$c_d(\nu, \mu) = \Theta(1) \cdot \left[ a^{1-\zeta} (1-a)^{(1-a)/a} [1 - (1-a)^d]^{\zeta/d} \right]^\nu. \quad (3.16)$$

Here  $0 < a < 1$  is the unique solution to the equation

$$1 - a = \exp [c(1 - (1-a)^{d-1})] = \exp \left[ -\zeta a \cdot \frac{1 - (1-a)^{d-1}}{1 - (1-a)^d} \right], \quad (3.17)$$

where the last expression is obtained by plugging in  $c = \zeta a/b$  and  $b = 1 - (1-a)^d$ . Now, (3.16) and (3.17) yield the desired formula for  $c_d(\nu, \mu)$ , so that we have established Theorem 1.

**Proof of Theorem 2.**

The proof of Theorem 2 relies on a similar idea as the proof of Theorem 1, but is a little simpler. Let  $\kappa(an)$  signify the probability that  $\mathcal{N}(H_d(n, p)) = an$ . We shall establish that

$$\kappa(an) \sim \binom{n}{an} c_d(an, p) (1-p)^{\binom{n}{d} - \binom{an}{d} - \binom{(1-a)n}{d}}. \quad (3.18)$$

Then, we will estimate  $\kappa(an)$  and solve (3.18) for  $c_d(an, p)$  to obtain the formula stated in Theorem 2.

To prove (3.18), note that the term on the right hand side is just the expected number of components of order  $an$  that occur in  $H_d(n, p)$ , and thus provides an upper bound on  $\kappa(an)$ . For there are  $\binom{n}{an}$  ways to choose a set  $S$  of  $an$  vertices where to place such a component, and as each edge contained in  $S$  is present with probability  $p$  independently,  $c_d(an, p)$  is the probability that the subhypergraph of  $H_d(n, p)$  induced on  $S$  is connected. Moreover, this subhypergraph is a component iff none of the  $\binom{n}{d} - \binom{an}{d} - \binom{(1-a)n}{d}$  possible edges connecting  $S$  with  $V \setminus S$  is present in  $H_d(n, p)$ , and each of these edges occurs with probability  $p$  independently. On the other hand, we have

$$\kappa(an) \geq \binom{n}{an} c_d(an, p) (1-p)^{\binom{n}{d} - \binom{an}{d} - \binom{(1-a)n}{d}} \mathbb{P}[\mathcal{N}(H_d((1-a)n, p)) < an], \quad (3.19)$$

because the term on the right hand side equals the probability that there is precisely one component of order  $an$ . As  $\mathbb{P}[\mathcal{N}(H_d((1-a)n, p)) < an] \sim 1$  by Lemma 2, (3.18) follows from (3.19).

The next proposition, which we shall prove in Section 3.5, yields the desired estimate on  $\kappa(an)$ .

**Proposition 2** *We have  $\kappa(an) = \Theta_c((1-a)n)^{-1/2}$ .*

Combining Proposition 2 and (3.18) and estimating  $\binom{n}{an}$  via Stirling's formula, we get

$$c_d(an, p) = \Theta_c(1) \cdot [a(1-a)^{(1-a)/a}]^{an} (1-p)^{\binom{(1-a)n}{d} + \binom{an}{d} - \binom{n}{d}}. \quad (3.20)$$

To further evaluate the term on the right hand side, we need the following auxiliary lemma, whose proof can be found in Section 3.4.4.

**Lemma 4** We have  $(1-p)^{\binom{1-a}{d}n} + \binom{an}{d} - \binom{n}{d} = \Theta_c(1) \cdot \exp\left[\frac{c}{d}(1-a^d - (1-a)^d)n\right]$ .

Combining Lemma 4 with (3.20) and setting  $\nu = an$  and  $\zeta = a^{d-1}c$ , we obtain

$$c_d(\nu, p) = \Theta_c(1) \cdot \left[ a(1-a)^{(1-a)/a} \exp\left(\frac{\zeta(1-a^d - (1-a)^d)}{da^d}\right) \right]^\nu,$$

where  $0 < a < 1$  is the solution of the equation  $1-a = \exp\left[c((1-a)^{d-1} - 1)\right] = \exp\left[\frac{\zeta((1-a)^{d-1} - 1)}{a^{d-1}}\right]$ .

Thus, we have established Theorem 2.

#### Proof of Theorem 4.

Consider the following simple procedure for sampling a connected hypergraph of order  $\nu$  and size  $\mu$ .

Let  $\zeta = d\mu/\nu$ , compute the solution  $0 < a < 1$  of (3.17), let  $b = 1 - (1-a)^d$ , and set  $n = a^{-1}\nu$ ,  $m = b^{-1}\mu$ . Then, sample a random hypergraph  $H_d(n, m)$  and determine its largest connected component  $\mathcal{C}$ . If  $\mathcal{C}$  has order  $\nu$  and size  $\mu$ , then output  $\mathcal{C}$ ; otherwise, sample another  $H_d(n, m)$  independently, and so on.

By Proposition 1, the expected number of samples of  $H_d(n, m)$  needed until  $\mathcal{C}$  has order  $\nu$  and size  $\mu$  is  $O(\sqrt{nm})$ . Since the largest component of  $H_d(n, m)$  can be determined in time  $O(m)$ , the expected running time of the sampling procedure is  $O(n^{1/2}m^{3/2}) = O(\nu^{1/2}\mu^{3/2})$ . Moreover, given that  $\mathcal{C}$  has order  $\nu$  and size  $\mu$ ,  $\mathcal{C}$  is a uniformly distributed connected hypergraph with these parameters. Finally, it is easily seen that given  $\nu$  and  $\mu$ , the solution  $0 < a < 1$  of (3.17) can be found in time  $O(\nu)$  (e.g., via binary search). Thus, we have established that this simple algorithm samples a uniformly distributed connected hypergraph of order  $\nu$  and size  $\mu$  in expected time  $O(\nu^{1/2}\mu^{3/2})$ .

### 3.4.2 Proof of Lemma 2

The expected average degree of  $H_d((1-a)n, p)$  is  $c' = \binom{(1-a)n-1}{d-1} p \sim (1-a)^{d-1} c$ . Therefore, by the first part of Theorem 5, it suffices to show that

$$c' \sim (1-a)^{d-1} c < (d-1)^{-1} \quad \text{if } c > (d-1)^{-1}. \quad (3.21)$$

Let  $x_0 = (c(d-1))^{-1/(d-1)}$ , and set  $f(x) = \exp(c(x^{d-1} - 1))$ . Then  $1-a$  is the unique fixed point of  $f$  in the interval  $(0, 1)$ . Moreover, since  $f(0) > 0$  and  $f(1) = 1$ , the function  $f$  lies above the identity  $x \mapsto x$  in the interval  $(0, 1-a)$ , and below  $x \mapsto x$  in the interval  $(1-a, 1)$ . Therefore, in order to prove (3.21) it suffices to show that  $f(x_0) < x_0$ , i.e., that  $c(d-1) \exp(1 - c(d-1)) < 1$  for all  $c > 1/(d-1)$ . Letting  $z = (d-1)c$ , we obtain the equivalent condition

$$h(z) = z \exp(1-z) < 1 \quad \text{for all } z > 1. \quad (3.22)$$

Since  $\frac{\partial}{\partial z} \ln h(z) = \frac{1}{z} - 1$ ,  $h$  is strictly decreasing for  $z > 1$ , so that (3.22) follows from the fact that  $h(1) = 1$ .

### 3.4.3 Proof of Lemma 3

To prove 1., we first compare  $a^d \binom{n}{d}$  and  $\binom{an}{d}$ :

$$\begin{aligned} a^d \binom{n}{d} \binom{an}{d}^{-1} &= \frac{a^d (n)_d}{(an)_d} = \prod_{j=0}^{d-1} \frac{a(n-j)}{an-j} = \prod_{j=0}^{d-1} 1 + \frac{(1-a)j}{an-j} \\ &\stackrel{(3.7)}{=} \exp \left[ \sum_{j=0}^{d-1} \frac{(1-a)j}{an} + O_c(n^{-2}) \right] = \exp \left[ \binom{d}{2} \frac{1-a}{an} + O_c(n^{-2}) \right]. \end{aligned} \quad (3.23)$$

Hence, as  $bm = o(n^2)$ , we obtain

$$\begin{aligned} \left[ \binom{an}{d} p (bm)^{-1} \right]^{bm} &= \left[ a^d \binom{n}{d} p (bm)^{-1} \right]^{bm} \exp \left[ - \binom{d}{2} \frac{(1-a)bm}{n} + o(1) \right] \\ &\sim \left( \frac{a^d}{b} \right)^{bm} \exp \left[ - \frac{(d-1)(1-a)bc}{2} \right]. \end{aligned} \quad (3.24)$$

Chapter 3. Counting Connected Graphs and Hypergraphs via the Probabilistic Method

As  $0 \leq b \leq 1$  and  $1 - a = \exp(-\Omega(c))$  by (3.3), the last factor in (3.24) is  $\Theta(1)$ , so that 1. follows.

With respect to 2., we have

$$\begin{aligned}
 \xi &= \left[ \left( \binom{an}{d} - bm \right) \binom{an}{d}^{-1} \right]^{(an)_d - bm} = \left[ 1 - bm \binom{an}{d}^{-1} \right]^{(an)_d - bm} \\
 &\stackrel{(3.7)}{\sim} \exp \left[ - \left( \binom{an}{d} - bm \right) \left( \frac{bm}{\binom{an}{d}} + \frac{b^2 m^2}{2 \binom{an}{d}^2} \right) \right] \\
 &= \exp \left[ -bm + \frac{b^2 m^2}{2 \binom{an}{d}} - \frac{b^3 m^3}{\binom{an}{d}^2} \right] \sim \exp \left[ -bm + \frac{b^2 m^2}{2 \binom{an}{d}} \right], \tag{3.25}
 \end{aligned}$$

because  $(bm)^3 \binom{an}{d}^{-2} = o(1)$ . Furthermore,

$$\begin{aligned}
 \eta &= (1-p)^{-\left[ \binom{n}{d} - \binom{(1-a)n}{d} - bm \right]} \stackrel{(3.7)}{\sim} \exp \left[ \left( p + \frac{p^2}{2} \right) \left( \binom{n}{d} - \binom{(1-a)n}{d} - bm \right) \right] \\
 &\sim \exp \left[ m - \binom{(1-a)n}{d} p - bmp + \frac{mp}{2} - \frac{p^2}{2} \binom{(1-a)n}{d} \right]. \tag{3.26}
 \end{aligned}$$

As  $(1-a)^{-1} = \exp(\Theta(c)) = n^{o(1)}$  by (3.3) and our assumption that  $c = o(\ln n)$ , we can compare  $(1-a)^d \binom{n}{d}$  and  $\binom{(1-a)n}{d}$  as follows:

$$\begin{aligned}
 \frac{\binom{(1-a)n}{d}}{(1-a)^d \binom{n}{d}} &= \frac{\binom{(1-a)n}{d}}{\binom{(1-a)n}{d} \binom{n}{d}} = \prod_{j=0}^{d-1} 1 - \frac{aj}{(1-a)(n-j)} \stackrel{(3.7)}{=} \exp \left[ - \sum_{j=0}^{d-1} \frac{aj}{(1-a)n} + O(n^{o(1)-2}) \right] \\
 &= \exp \left[ - \binom{d}{2} \frac{a}{(1-a)n} + O(n^{o(1)-2}) \right] = 1 - \binom{d}{2} \frac{a}{(1-a)n} + O(n^{o(1)-2}). \tag{3.27}
 \end{aligned}$$

Plugging (3.27) into (3.26) and recalling that  $b = 1 - (1-a)^d$ , we get

$$\begin{aligned}
 \eta &\sim \exp \left[ (1 - (1-a)^d)m + \binom{d}{2} \frac{a(1-a)^d m}{(1-a)n} + \frac{mp}{2} (1 - (1-a)^d - 2b) \right] \\
 &= \exp \left[ bm + \binom{d}{2} \frac{a(1-a)^{d-1} m}{n} - \frac{bmp}{2} \right]. \tag{3.28}
 \end{aligned}$$

Combining (3.25) and (3.28), we obtain

$$\begin{aligned} \xi \cdot \eta &\sim \exp \left[ \binom{d}{2} \frac{a(1-a)^{d-1}m}{n} - \frac{bmp}{2} + \frac{b^2m^2}{2\binom{an}{d}} \right] \stackrel{(3.23)}{\sim} \exp \left[ \binom{d}{2} \frac{a(1-a)^{d-1}m}{n} + \frac{bm}{2} \left[ \frac{bm}{a^d \binom{n}{d}} - p \right] \right] \\ &= \exp \left[ \frac{(d-1)a(1-a)^{d-1}c}{2} + \frac{bmp}{2} \cdot \frac{1-a^d - (1-a)^d}{a^d} \right]. \end{aligned} \quad (3.29)$$

By (3.3), we have

$$(d-1)a(1-a)^{d-1}c = O(1). \quad (3.30)$$

Further, as  $bmp = O(n^{2-d}c^2)$ , we have  $bmp = o(1)$  if  $d > 2$ . Moreover, if  $d = 2$ , then (3.3) entails that

$$bmp \cdot \frac{1-a^2 - (1-a)^2}{a^2} = O_c(c^2) \cdot \frac{2(1-a)}{a} = O_c[c^2 \cdot \exp(-\Omega(c))] = O_c(1). \quad (3.31)$$

Thus, plugging (3.30) and (3.31) into (3.29), we get  $\xi \cdot \eta = \Theta_c(1)$ , thereby completing the proof.

### 3.4.4 Proof of Lemma 4

By (3.23) and (3.27), we have

$$\begin{aligned} (1-p)^{\binom{(1-a)n}{d} + \binom{an}{d} - \binom{n}{d}} &\stackrel{(3.7)}{\sim} \exp \left[ \left( p + \frac{p^2}{2} \right) \left[ \binom{n}{d} - \binom{(1-a)n}{d} - \binom{an}{d} \right] \right] \\ &\sim \exp \left[ \binom{n}{d} p [1 - a^d - (1-a)^d] \right] \times \exp \left[ \frac{p^2}{2} \binom{n}{d} [1 - a^d - (1-a)^d] \right] \\ &\quad \times \exp \left[ \binom{n}{d} \binom{d}{2} n^{-1} \left( p + \frac{p^2}{2} \right) [a(1-a)^{d-1} + (1-a)a^{d-1}] \right]. \end{aligned} \quad (3.32)$$

With respect to the third factor, (3.3) entails that

$$\binom{n}{d} \binom{d}{2} n^{-1} \left( p + \frac{p^2}{2} \right) [a(1-a)^{d-1} + (1-a)a^{d-1}] = O(c(1-a)) = O(1). \quad (3.33)$$

Furthermore, as  $\binom{n}{d}p^d = o(1)$  if  $d \geq 2$ , we only need to consider the second factor in (3.32) for  $d = 2$ . In this case, we have  $1 - a^2 - (1 - a)^2 = 2(1 - a)$ , so that (3.3) implies

$$p^2 \binom{n}{d} (1 - a) = O(c^2(1 - a)) = O(1). \quad (3.34)$$

Plugging (3.33) and (3.34) into (3.32), we obtain the desired estimate.

## 3.5 The Probability of Getting a Giant Component of a Given Order and Size

### 3.5.1 Outline

Throughout Section 3.5, we assume that  $c_0 \leq c = o(\ln n)$  for some constant  $c_0 > (d - 1)^{-1}$ . We set  $m = cn/d$  and let  $a, b$  be as in Theorem 5. The goal of this section is to prove Propositions 1 and 2. For instance, Proposition 1 claims that the probability that both the order and the size of the largest component of  $H_d(n, p)$  equal *exactly* their expectations  $an$  and  $bm$  is  $\Theta_c((1 - a)bm n)^{-1/2}$ . By comparison, Theorem 5 states that the variance of  $\mathcal{N}(H_d(n, p))$  is  $O_c((1 - a)n)$ , so that by Chebyshev's inequality  $\mathcal{N}(H_d(n, p))$  is concentrated in width  $O_c(\sqrt{(1 - a)n})$  about  $an$ . Similarly,  $\mathcal{M}(H_d(n, p))$  is concentrated in width  $O_c(\sqrt{bm})$  about  $bm$ . Therefore, it seems reasonable that indeed  $\chi(an, bm) = \Theta_c((1 - a)bm n)^{-1/2}$ . However, this estimate cannot be derived directly from Theorem 5, because there is no *a priori* reason why  $\mathcal{N}(H_d(n, p))$  (or  $\mathcal{M}(H_d(n, p))$ ) should be “approximately uniformly” distributed in the interval in which it is concentrated.

In order to prove Propositions 1 and 2, we expose the edges of the random hypergraph  $H_d(n, p)$  in two rounds. In the first round, we choose a random hypergraph  $H_1 = H_d(n, p_1)$ , where  $p_1 = (1 - \varepsilon)p$  for some sufficiently small constant  $\varepsilon > 0$ . Then, in the second round we select every edge that is not present in  $H_1$  with probability  $p_2$  independently, where  $p_2$  is defined by the equation  $p_1 + p_2 - p_1 p_2 = p$ . Let  $F$  signify the set of all edges selected the second

round. By the choice of  $p_1$  and  $p_2$ , in the hypergraph  $H_1 + F$  each possible edge is present with probability  $p$  independently, so that  $H_1 + F = H_d(n, p)$ .

Choosing  $\varepsilon = \varepsilon(c_0)$  sufficiently small, we can ensure that  $\binom{n-1}{d-1}p_1 > (d-1)^{-1} + \Omega(1)$ . Hence, by Theorem 5  $H_1$  has a unique giant component  $\mathcal{C}$  w.h.p. Let  $N = |V(\mathcal{C})|$  and  $M = |E(\mathcal{C})|$ . When we add the edges  $F$  to  $H_1$ , further vertices and edges get attached to the component  $\mathcal{C}$ . To estimate the probability that  $\mathcal{N}(H_1 + F)$  equals  $an$ , we study the number of vertices in components of order  $\geq 2$  of  $H_1$  that get attached to  $\mathcal{C}$  via  $F$ . In addition, we show that the number of isolated vertices of  $H_1$  that get connected with  $\mathcal{C}$  via  $F$  is binomially distributed. Then, we can apply Lemma 1 to estimate the probability that  $\mathcal{N}(H_1 + F) = an$ . Furthermore, to study  $\mathcal{M}(H_1 + F)$ , we observe that the number of edges in  $F$  that lie completely inside of  $\mathcal{C}$  is binomially distributed, so that we can apply Lemma 1 once more to get the probability that also  $\mathcal{M}(H_1 + F) = bm$ .

Let us now implement this approach in detail. In order to keep track of the growth of  $\mathcal{C}$ , we do not add all the edges in  $F$  at once, but we partition  $F$  into five sets  $F_0, \dots, F_4$  and add these sets one by one. First, we let  $F_0 = \{e \in F : e \subset V(\mathcal{C})\}$ ,  $F_1 = \{e \in F : e \subset V \setminus V(\mathcal{C})\}$ , and  $H_2 = H_1 + F_1$ ; that is,  $H_2$  is obtained from  $H_1$  by adding all edges in  $F$  that lie completely outside of  $\mathcal{C}$ . Hence,  $\mathcal{C}$  is still a connected component of  $H_2$ . Observe that

$$|F_0| = \text{Bin} \left[ \binom{N}{d} - M, p_2 \right]. \quad (3.35)$$

The following lemma bounds the mean of  $|F_0|$ .

**Lemma 5** *There is a number  $K_0 = \Omega_c(1)$  such that  $\mathbb{P} \left[ \binom{N}{d} - M \geq K_0 \binom{n}{d} \right] \geq 1 - n^{-9}$ .*

**Proof** Theorem 5 shows that with probability  $\geq 1 - n^{-10}$  we have  $N = \Omega_c(n)$ . Furthermore,  $M$  is bounded from above by the total number of edges of  $H_d(n, p_1)$ , which is binomially distributed with mean  $\binom{n}{d}p_1 = O(n \ln n)$ . Hence, the Chernoff bound (3.8) entails that  $M = o(n^d)$  with probability  $\geq 1 - n^{-10}$ . Hence, with probability  $\geq 1 - n^{-9}$  we have  $\binom{N}{d} - M = \binom{\Omega_c(n)}{d} - o(n^d) = \Omega_c \binom{n}{d}$ .  $\square$

Furthermore, let  $F_2$  be the set of all  $e \in F$  such that either  $1 \leq |e \cap V(\mathcal{C})| < d - 1$ , or  $e$  connects  $\mathcal{C}$  with a component of  $H_2 - \mathcal{C}$  of order  $\geq 2$ . In addition, let  $F_3$  be the set of all  $e \in F \setminus (F_1 \cup F_2)$  such that there is an  $e' \in F_2$  satisfying  $e \cap e' \setminus V(\mathcal{C}) \neq \emptyset$ . Then  $F_3$  consists of all edges  $e$  in  $F \setminus F_2$  that connect an isolated vertex  $v$  of  $H_2$  with  $\mathcal{C}$  so that  $v$  is also connected with  $\mathcal{C}$  via an edge in  $F_2$ . Let  $H_3 = H_2 + F_2 + F_3$ . Moreover, let  $\mathcal{C}'$  be the component of  $H_3$  that contains  $\mathcal{C}$ , and set  $N' = |V(\mathcal{C}')|$ ,  $M' = |E(\mathcal{C}')|$ .

Finally, let  $F_4 = \{e \in F : |e \cap V(\mathcal{C})| = d - 1 \text{ and } e \text{ contains an isolated vertex of } H_3\}$ . Let  $Y$  denote the number of isolated vertices of  $H_3$ , and let  $Z$  be the total number of isolated vertices that the edges in  $F_4$  connect with  $\mathcal{C}'$ . Clearly,  $Z \leq |F_4|$ . Moreover, observe that each isolated vertex  $v$  of  $H_3$  is contained in  $\binom{N}{d-1}$  possible edges that have  $d - 1$  vertices in  $\mathcal{C}$ . As each of these  $\binom{N}{d-1}$  edges is present in  $F$  with probability  $p_2$  independently, the probability that  $v$  gets attached to  $\mathcal{C}$  via an edge in  $F_4$  is

$$r = 1 - (1 - p_2)^{\binom{N}{d-1}}, \text{ so that } Z = \text{Bin}[Y, r], \text{ and } |F_4| = \text{Bin}\left[Y \cdot \binom{N}{d-1}, p_2\right]. \quad (3.36)$$

Let  $H_4 = H_3 + F_0 + F_4$ , let  $\mathcal{C}''$  denote the component of  $H_4$  that contains  $\mathcal{C}$ , and let  $N'' = |V(\mathcal{C}'')|$ ,  $M'' = |E(\mathcal{C}'')|$ . Then

$$N'' = N' + Z, \quad M'' = |F_0| + |F_4| + M'. \quad (3.37)$$

Moreover, as  $F = \bigcup_{i=0}^4 F_i$ , we have  $H_4 = H_1 + F$ , so that  $H_4$  is distributed as  $H_d(n, p)$ .

Let us point out that  $\mathcal{C}''$  is the unique component of  $H_4$  that has order  $\Omega_c(n)$  w.h.p.

**Lemma 6** *With probability  $\geq 1 - n^{-9}$  we have  $N'' = \mathcal{N}(H_4)$  and  $M'' = \mathcal{M}(H_4)$ . Moreover, there exists a number  $K_1 = O_c(1)$  such that*

$$\alpha_0 = \mathbb{P}\left[|N'' - an| \geq \frac{K_1}{2} \sqrt{(1-a)n}\right] \leq 0.001. \quad (3.38)$$

$$\beta_0 = \mathbb{P}\left[|M'' - bm| \geq \frac{K_1}{3} n^{d/2} p^{1/2}\right] \leq 0.001. \quad (3.39)$$

**Proof** By Theorem 5, with probability  $\geq 1 - n^{-10}$   $\mathcal{C}$  is the unique component of order  $\Omega_c(n)$  of  $H_1 = H_d(n, p_1)$ , because  $\binom{n-1}{d-1}p_1$  is bounded away from  $(d-1)^{-1}$  by our choice of  $\varepsilon$ . Thus, as  $\mathcal{C}$  is contained in  $\mathcal{C}''$ , with probability  $\geq 1 - n^{-10}$  the component  $\mathcal{C}''$  of  $H_4$  has order  $\Omega_c(n)$ . Furthermore, since  $H_4$  is distributed as  $H_d(n, p)$ , Theorem 5 entails that with probability  $\geq 1 - n^{-10}$   $H_4$  has precisely one component of order  $\Omega_c(n)$ , so that  $N'' = |V(\mathcal{C}'')| = \mathcal{N}(H_4)$  and  $M'' = |E(\mathcal{C}'')| = \mathcal{M}(H_4)$ . Hence, with probability  $\geq 1 - 2n^{-10} \geq 1 - n^{-9}$  we have  $N'' = \mathcal{N}(H_4)$ ,  $M'' = \mathcal{M}(H_4)$ .

As a consequence, the estimate of the variance of  $\mathcal{N}(H_d(n, p))$  given in Theorem 5 applies to the random variable  $N''$ . Note that the expression on the right hand side of (3.5) is  $O_c((1-a)n)$ . Hence,  $\text{Var}(N'') = O_c((1-a)n)$ , so that the bound on  $\alpha_0$  follows from Chebyshev's inequality (provided that  $K_1$  is chosen large enough). Moreover, the bound on  $\beta_0$  is a direct consequence of (3.6).  $\square$

Lemma 6 implies that in order to compute  $\chi(an, bm)$ , we just need to estimate the probability that  $N'' = an$  and  $M'' = bm$ . To achieve this goal, we shall first derive the probable value of  $Y$ , thereby determining the distribution of  $|F_4|$  and  $Z$  (cf. Lemma 7 below). Then, we use Theorem 5 to estimate  $N'$  and  $M'$  (cf. Lemma 8). Finally, since  $Z$  and  $|F_0|$  are binomially distributed by (3.36), we can apply Lemma 1 in order to estimate the probability that  $Z = an - N'$  and  $|F_0| = bm - |F_4| - M'$  (cf. Lemma 9); by (3.37), this yields precisely the probability that  $N'' = an$  and  $M'' = bm$ .

With respect to the probable value of  $Y$ , we shall prove the following in Section 3.5.2.

**Lemma 7** *There exist constants  $K_2, K_3 > 0$  such that  $\text{P} [K_2(1-a)n \leq rY \leq K_3(1-a)n] \geq 1 - \exp(-n^{\Omega(1)})$ .*

To determine the relation between  $Y$ ,  $|F_0|$ ,  $|F_4|$ ,  $N'$ , and  $M'$ , let

$$\Delta = |N' + rY - an|, \quad \Gamma = \left| M' + Y \binom{N}{d-1} p_2 + \left[ \binom{N}{d} - M \right] p_2 - bm \right|. \quad (3.40)$$

Recall that  $Z$  is binomially distributed with mean  $rY$  (cf. (3.36)). Hence, if  $\Delta$  is “small” – say, of the same order of magnitude as the standard deviation  $\sqrt{r(1-r)Y}$  of  $Z$  – then we can apply Lemma 1 to compute the probability that  $Z$  attains precisely the value  $an - N'$ , so that  $N'' = N' + Z = an$ . Similarly, if  $\Gamma$  is “small”, then Lemma 1 yields an estimate of the probability that  $|F_0| = bm - M' - |F_4|$ , i.e.,  $M'' = |F_0| + |F_4| + M' = bm$ . The next lemma shows that it is in fact rather likely that  $\Gamma$  and  $\Delta$  are small. We shall prove Lemma 8 in Section 3.5.3.

**Lemma 8** *There is a number  $K_4 = O_c(1)$  so that  $\mathbb{P} [\Delta \leq K_4((1-a)n)^{1/2}, \Gamma \leq K_4n^{d/2}p^{1/2}] \geq 0.99$ .*

Furthermore, in Section 3.5.4 we combine Lemmas 1, 7 and 8 to prove the following lower bound on the probability that  $Z = an - N'$  and  $|F_0| = bm - |F_4| - M'$ .

**Lemma 9** *Let  $\mathcal{E}$  be the event that the following conditions are satisfied:*

**E1.**  $\mathcal{C}''$  is the unique component of  $H_1 + F$  of order  $\Omega_c(n)$ ,

**E2.**  $K_2(1-a)n \leq rY \leq K_3(1-a)n$ .

**E3.**  $\binom{N}{d} - M \geq K_0 \binom{n}{d}$ ,

**E4.**  $\Delta \leq K_4\sqrt{(1-a)n}$  and  $\Gamma \leq K_4n^{d/2}p^{1/2}$ .

Then

$$\begin{aligned} \mathbb{P} [Z = an - N' \wedge |F_0| = bm - |F_4| - M' | \mathcal{E}] &= \Omega_c \left( [(1-a)bnm]^{-1/2} \right), \\ \mathbb{P} [Z = an - N' | \mathcal{E}] &= \Omega_c \left( [(1-a)n]^{-1/2} \right). \end{aligned}$$

*Proof of Propositions 1 and 2.* Combining Lemmas 5–8, we conclude that  $\mathbb{P} [\mathcal{E}] \geq 0.98$ . Thus, Lemma 9 and (3.37) entail that

$$\kappa(an) \geq \mathbb{P} [Z = an - N' | \mathcal{E}] \mathbb{P} [\mathcal{E}] = \Omega_c \left( [(1-a)n]^{-\frac{1}{2}} \right). \quad (3.41)$$

$$\chi(an, bm) \geq \mathbb{P}[Z = an - N' \wedge |F_0| = bm - |F_4| - M' | \mathcal{E}] \mathbb{P}[\mathcal{E}] = \Omega_c \left[ [(1-a)bnm]^{-\frac{1}{2}} \right]. \quad (3.42)$$

On the other hand, (3.9) yields in combination with (3.36) that

$$\mathbb{P}[Z = an - N' | K_2(1-a)n \leq rY] = O \left( [(1-a)n]^{-1/2} \right). \quad (3.43)$$

Similarly, as  $|F_0|$  and  $Z$  are independent, combining (3.9) and (3.35), we obtain

$$\mathbb{P} \left[ |F_0| = bm - |F_4| - M' | Z = an - N', \binom{N}{d} - M \geq K_0 \binom{n}{d} \right] = O_c \left[ [bm]^{-\frac{1}{2}} \right]. \quad (3.44)$$

Hence, invoking Lemmas 5 and 7, we get

$$\begin{aligned} \chi(an, bm) &\stackrel{(3.37)}{=} \mathbb{P}[|F_0| = bm - |F_4| - M' | Z = an - N'] \mathbb{P}[Z = an - N'] \\ &\stackrel{(3.43), (3.44)}{\leq} O_c \left( [(1-a)bnm]^{-1/2} \right) + \mathbb{P}[Y < K_2(1-a)n] + \mathbb{P} \left[ \binom{N}{d} - M < K_0 \binom{n}{d} \right] \\ &= O_c \left( [(1-a)bnm]^{-1/2} \right). \end{aligned} \quad (3.45)$$

Finally, combining (3.42) and (3.45), we obtain  $\chi(an, bm) = \Theta_c((1-a)bnm)^{-1/2}$ . Similarly, combining (3.41) and (3.43), we get  $\kappa(an) = \Theta_c((1-a)n)^{-1/2}$ .

### 3.5.2 Proof of Lemma 7

Define  $\mathcal{Y} = \{y : |(1-r)y - n \exp(-c)| > 2n^{9/10}\}$ . Then (3.3) and the definition  $1-a = \exp(c((1-a)^{d-1} - 1))$  of  $a$  yield  $\frac{1-a}{\exp(-c)} = \exp(c[(1-a)^{d-1}]) = \Theta(1)$ . Hence, if  $Y \notin \mathcal{Y}$ , then  $rY = \Theta((1-a)n)$ .

To bound the probability that  $Y \in \mathcal{Y}$ , let  $I = Y - Z$  signify the number of isolated vertices of  $H_1 + F$ . Then on the one hand

$$\begin{aligned} \mathbb{P}[|I - n \exp(-c)| \leq n^{9/10} | Y \in \mathcal{Y}] &\leq \mathbb{P}[|Z - rY| > n^{9/10} | Y \in \mathcal{Y}] \\ &\stackrel{(3.36)}{=} \mathbb{P}[|\text{Bin}[Y, r] - rY| > n^{9/10} | Y \in \mathcal{Y}] \stackrel{(3.8)}{=} o(1). \end{aligned} \quad (3.46)$$

On the other hand, as  $H_1 + F = H_d(n, p)$ , the statement on the number of isolated vertices of  $H_d(n, p)$  in Theorem 5 shows that  $\mathbb{P}[|I - n \exp(-c)| \leq n^{9/10}] = 1 - \exp(-n^{\Omega(1)})$ . Therefore, (3.46) implies that  $\mathbb{P}[Y \in \mathcal{Y}] = \exp[-n^{\Omega(1)}]$ .

### 3.5.3 Proof of Lemma 8

In order to bound  $\Delta = |N' + rY - an|$  we have to show that  $N'$  is concentrated about  $an - rY$ . Basically this follows from the fact that  $N'' = N' + Z$  is concentrated about  $an$  by Lemma 6 and that  $Z = \text{Bin}(Y, r)$  is concentrated about  $rY$  by Chernoff bounds. To give a precise argument, we let  $C \geq K_1$  be a sufficiently large number such that (3.38) in Lemma 6 is satisfied. Let  $\alpha_0$  be as in (3.38). By Lemma 7, there is a constant  $K_3 > 0$  such that

$$\mathbb{P}[rY \leq K_3(1-a)n] = 1 - o(1). \quad (3.47)$$

Since by (3.36)  $Z$  is binomially distributed with mean  $rY$ , we obtain

$$\begin{aligned} \alpha_1 &= \mathbb{P}\left[|Z - rY| \geq \frac{C}{2}\sqrt{(1-a)n}\right] \leq \mathbb{P}\left[|Z - rY| \geq \frac{C}{2}\sqrt{(1-a)n} \mid rY \leq C'(1-a)n\right] \\ &\quad + \mathbb{P}[Y > K_3(1-a)r^{-1}n] \stackrel{(3.8), (3.47)}{\leq} 0.001, \end{aligned} \quad (3.48)$$

provided that  $C$  is chosen large enough. As  $N'' = N' + Z$ , combining (3.38) and (3.48) yields

$$\mathbb{P}\left[\Delta \geq C\sqrt{(1-a)n}\right] \leq \alpha_0 + \alpha_1 \leq 0.002. \quad (3.49)$$

To bound the probability that  $\Gamma = |M' + Y\binom{N}{d-1}p_2 + (\binom{N}{d} - M)p_2 - bm|$  is large, we have to show that  $M'$  is concentrated about  $bm - Y\binom{N}{d-1}p_2 - (\binom{N}{d} - M)p_2$ . This essentially follows from the fact that  $M'' = M' + |F_0| + |F_4|$  is concentrated about  $bm$  by Lemma 6, while the binomially distributed variables  $|F_0|, |F_4|$  are concentrated about their means  $(\binom{N}{d} - M)p_2$  and  $Y\binom{N}{d-1}p_2$  by Chernoff bounds. To be precise, let  $\beta_0$  be as in (3.38). As  $|F_0|$  and  $|F_4|$  are binomially distributed

by (3.35) and (3.36), the Chernoff bound (3.8) yields

$$\beta_1 = \mathbb{P} \left[ \left| |F_0| - \left[ \binom{N}{d} - M \right] p_2 \right| \geq \frac{C}{3} n^{d/2} p^{1/2} \right] \leq 0.001, \quad (3.50)$$

$$\beta_2 = \mathbb{P} \left[ \left| |F_4| - Y \binom{N}{d-1} p_2 \right| \geq \frac{C}{3} n^{d/2} p^{1/2} \right] \leq 0.001, \quad (3.51)$$

provided that  $C$  is large enough. As  $M'' = M' + |F_0| + |F_4|$ , the definition (3.40) of  $\Gamma$  and (3.39), (3.50), and (3.51) entail that

$$\begin{aligned} \mathbb{P} [\Gamma \geq C n^{d/2} p^{1/2}] &= \mathbb{P} \left[ \left| M' + Y \binom{N}{d-1} p_2 + \left[ \binom{N}{d} - M \right] p_2 - bm \right| \geq C n^{d/2} p^{1/2} \right] \\ &\leq \beta_0 + \beta_1 + \beta_2 \leq 0.003. \end{aligned} \quad (3.52)$$

Thus, the assertion follows from (3.49) and (3.52).

### 3.5.4 Proof of Lemma 9

By condition E2 and (3.36), given that  $\mathcal{E}$  occurs  $Z$  is binomially distributed with mean  $rY = \Omega((1-a)n)$ . Furthermore, by E4 the desired value  $an - N'$  of  $Z$  satisfies  $|an - N' - rY| = \Delta \leq K_4 \sqrt{(1-a)n}$ , where  $K_4 = O_c(1)$ . In other words,  $an - N'$  is “close” to the expectation  $rY$  of  $Z$ . Hence, as the variance of  $Z$  is  $(1-r)rY = \Omega((1-a)n)$ , Lemma 1 entails that

$$\mathbb{P} [N'' = an \mid \mathcal{E}] \stackrel{(3.37)}{=} \mathbb{P} [Z = an - N' \mid \mathcal{E}] = \Omega_c \left( [(1-a)n]^{-1/2} \right). \quad (3.53)$$

In order to estimate  $\mathbb{P} [M'' = bm \mid \mathcal{E}]$ , we need the following observation.

**Lemma 10** *Let  $z = rY + O(\sqrt{(1-a)n})$ . Then  $\mathbb{P} [|F_4| = Y \binom{N}{d-1} p_2 + O(\sqrt{n}) \mid Z = z, \mathcal{E}] \geq 0.99$ .*

**Proof** Suppose that  $Z = z$ , and let  $v_1, \dots, v_z$  signify the isolated vertices of  $H_3$  that get connected with  $\mathcal{C}$  via the edges in  $F_4$ . Then the number of edges in  $F_4$  that are incident with  $v_i$

is distributed as  $\text{Bin}_{\geq 1}\left(\binom{N}{d-1}, p_2\right)$  (cf. Section 3.3), and these numbers are mutually independent for  $i = 1, \dots, z$ . Hence,

$$\begin{aligned} \mathbb{E}(|F_4| \mid Z = z) &= z \cdot \mathbb{E} \left[ \text{Bin}_{\geq 1} \left( \binom{N}{d-1}, p_2 \right) \right] \stackrel{(3.10)}{=} \frac{z}{r} \binom{N}{d-1} p_2 \\ &= Y \binom{N}{d-1} p_2 + O(\sqrt{(1-a)c^2 n}) \stackrel{(3.3)}{=} Y \binom{N}{d-1} p_2 + O(\sqrt{n}), \\ \text{Var}(|F_4| \mid Z = z) &= z \cdot \text{Var} \left[ \text{Bin}_{\geq 1} \left( \binom{N}{d-1}, p_2 \right) \right] \stackrel{(3.10)}{=} O(cz) \stackrel{\text{E2}}{=} O((1-a)cn) \stackrel{(3.3)}{=} O(n). \end{aligned}$$

Thus, the assertion follows from Chebyshev's inequality.  $\square$

Now, set  $z = an - N'$ . Let  $\mathcal{E}'$  be the following event:

$$\mathcal{E} \text{ occurs, } Z = z, \text{ and } |F_4| = Y \binom{N}{d-1} p_2 + O(\sqrt{n}). \quad (3.54)$$

We are going to estimate the probability that  $|F_0| = \varphi = bm - M' - |F_4|$ , given that  $\mathcal{E}'$  occurs. If  $\mathcal{E}'$  occurs, then by the definition (3.40) of  $\Gamma$  we have

$$\left| \varphi - \left[ \binom{N}{d} - M \right] p_2 \right| \leq \Gamma + \left| Y \binom{N}{d-1} p_2 - |F_4| \right| \stackrel{\text{E4}, (3.54)}{\leq} O(n^{d/2} p^{1/2}). \quad (3.55)$$

Therefore, again the desired value  $\varphi$  of  $|F_0|$  is ‘‘close’’ to the mean  $\left(\binom{N}{d} - M\right)p_2$  of  $|F_0|$ . Since  $|F_0|$  is binomially distributed with mean  $\left(\binom{N}{d} - M\right)p_2$  and variance  $(1 + o(1))\left(\binom{N}{d} - M\right)p_2$  by (3.35), and because given  $\mathcal{E}$  we have  $\left(\binom{N}{d} - M\right)p_2 = \Omega_c(n^d p)$ , Lemma 1 and (3.55) entail that

$$\mathbb{P}[|F_0| = \varphi \mid \mathcal{E}'] = \Omega_c(n^{-d/2} p^{-1/2}) = \Omega_c([bm]^{-1/2}). \quad (3.56)$$

Finally, Lemma 10 yields  $\mathbb{P}[Z = z \wedge |F_0| = \varphi \mid \mathcal{E}] = \mathbb{P}[Z = z \mid \mathcal{E}] \cdot \mathbb{P}[M'' = bm \mid \mathcal{E}, Z = z] \geq 0.99 \cdot \mathbb{P}[N'' = an \mid \mathcal{E}] \cdot \mathbb{P}[|F_0| = \varphi \mid \mathcal{E}']$ , so that the assertion follows from (3.53) and (3.56).

### 3.6 The Expected Number of Edges Given that $H_d(n, p)$ is Connected

In this section, we prove Theorem 3. Let  $c_0 > (d - 1)^{-1}$  be a constant, and suppose that  $c_0 \leq c = o(\ln n)$ . Let  $p = c/\binom{n-1}{d-1}$ . For a number  $0 \leq m \leq \binom{n}{d}$  we let

$$y(m) = \mathbb{P}[|E(H_d(n, p))| = m] = \binom{\binom{n}{d}}{m} p^m (1-p)^{\binom{n}{d}-m}.$$

Then the expected number of edges given that  $H_d(n, p)$  is connected is  $T = c_d(n, p)^{-1} \cdot \sum_{m=0}^{\binom{n}{d}} m \cdot y(m) c_d(n, m)$ . The term  $y(m) c_d(n, m) = \exp(-\Theta(n))$  is exponentially small, and we are interested in the behavior of the exponent, i.e., the number hidden in the  $\Theta$ -sign. To determine this number, we let  $m = \zeta n/d$  and consider

$$f(\zeta) = \lim_{n \rightarrow \infty} n^{-1} \ln y(\zeta n/d) = d^{-1} [\zeta - c + \zeta \ln(c/\zeta)] \quad (3.57)$$

(the last equation is due to Stirling's formula). Further, recall that  $\lim_{n \rightarrow \infty} n^{-1} \ln c_d(n, \zeta n/d) = \ln \Phi_d(\zeta)$ , where  $\Phi_d$  is the function defined in Theorem 1. Thus,

$$h(\zeta) = \lim_{n \rightarrow \infty} n^{-1} \ln [d^{-1} \zeta n \cdot y(m) c_d(n, m)] = f(\zeta) + \ln \Phi_d(\zeta).$$

**Lemma 11** *The function  $h(\zeta)$  has a unique global maximum at the point  $\zeta_0 = c\alpha_0^{-d}(1 - (1 - \alpha_0)^d)$ , where  $0 < \alpha_0 = \alpha_0(c) < 1$  is the solution of (3.2).*

**Proof** The function  $\ln \Phi_d(\zeta)$  increases strictly to 0 as  $\zeta \rightarrow \infty$ , whereas  $f(\zeta)$  is peaked around  $\zeta = c$  and  $f(\zeta) \rightarrow -\infty$  as  $\zeta \rightarrow \infty$ . Therefore, all global maxima of the function  $h(\zeta)$  lie in a compact interval  $\zeta' = c - \varepsilon \leq \zeta \leq \zeta''$ , where  $\varepsilon > 0$  is a sufficiently small constant. As we shall prove below that the derivative  $\frac{\partial}{\partial \zeta} h(\zeta)$  has a unique zero  $\zeta' \leq \zeta_0 \leq \zeta''$ , this  $\zeta_0$  is the unique global maximum of  $h(\zeta)$ .

Thus, let us compute  $\frac{\partial}{\partial \zeta} h(\zeta) = \frac{\partial}{\partial \zeta} f(\zeta) + \frac{\partial}{\partial \zeta} \ln \Phi_d(\zeta)$ . To compute  $\frac{\partial}{\partial \zeta} \ln \Phi_d(\zeta)$ , let  $\alpha = \alpha(\zeta)$  be the unique solution of the equation  $1 - \alpha = \exp(-\zeta \alpha \cdot \frac{1-(1-\alpha)^{d-1}}{1-(1-\alpha)^d})$  in the interval  $(0, 1)$

(cf. Theorem 1). Then

$$\begin{aligned} \zeta &= \zeta(\alpha) = -\frac{\ln(1-\alpha)}{\alpha} \cdot \frac{1-(1-\alpha)^d}{1-(1-\alpha)^{d-1}}, \quad \text{so that} \\ \frac{\partial \zeta}{\partial \alpha} &= \frac{1-(1-\alpha)^d}{\alpha(1-\alpha-(1-\alpha)^d)} + \frac{(1-(1-\alpha)^d)^2 + \alpha(\alpha-2+(1-\alpha)^d((d-1)\alpha+2))}{\alpha^2(1-\alpha-(1-\alpha)^d)^2} \ln(1-\alpha). \end{aligned} \quad (3.58)$$

Thus, using that  $\frac{\partial \alpha}{\partial \zeta} = \left[\frac{\partial \zeta}{\partial \alpha}\right]^{-1}$  and performing a tedious computation, we obtain

$$\frac{\partial}{\partial \zeta} \ln \Phi_d(\zeta) = d^{-1} \ln(1-(1-\alpha)^d) - \ln \alpha, \quad (3.59)$$

$$\frac{\partial}{\partial \zeta} f(\zeta) = d^{-1} \ln(c/\zeta) \stackrel{(3.58)}{=} d^{-1} \ln \left[ -\frac{c\alpha(1-(1-\alpha)^{d-1})}{(1-(1-\alpha)^d) \ln(1-\alpha)} \right]. \quad (3.60)$$

Hence, if  $\zeta_0$  is a zero of  $\frac{\partial}{\partial \zeta} h(\zeta) = \frac{\partial}{\partial \zeta} \ln \Phi_d(\zeta) + \frac{\partial}{\partial \zeta} f(\zeta)$ , then combining (3.59) and (3.60) and solving for  $\alpha$  in terms of  $c$ , we obtain that  $0 < \alpha_0 = \alpha(\zeta_0) < 1$  satisfies (3.2). Therefore, as the solution  $\alpha_0$  of (3.2) in the interval  $(0, 1)$  is unique, we conclude that  $\frac{\partial}{\partial \zeta} h(\zeta)$  has a unique zero  $\zeta_0 = \zeta(\alpha_0)$ . Finally, simplifying (3.58) using the relation (3.2) satisfied by  $\alpha_0$  yields the formula for  $\zeta_0$  stated in Lemma 11.  $\square$

Lemma 11 implies that  $T \sim \zeta_0 n/d$ . For as the global maximum  $\zeta_0$  is unique, for any fixed number  $\delta > 0$  we have

$$\sum_{|m-\zeta_0 n/d|>\delta} m \cdot y(m) c_d(n, m) \leq \exp(-\Omega(\delta n)) \cdot \sum_{|m-\zeta_0 n/d|\leq\delta} m \cdot y(m) c_d(n, m).$$

In other words, the only values of  $m$  that contribute to  $T$  are those that satisfy  $m \sim \zeta_0 n/d$ . Thus, we have established the formula stated in Theorem 3.

### 3.7 Branching Processes and The Giant Component of $H_d(n, p)$

In this section we prove Theorem 5 by establishing an analogy between a Galton-Watson branching process with a suitable successor distribution and exploring the components of a random

hypergraph. First we introduce the branching process in Section 3.7.1. In Section 3.7.2 prove the first part of Theorem 5 and investigate the expected order and size of the giant component for  $c > (d-1)^{-1}$ . Then, in Section 3.7.3 we show that indeed with probability  $\geq 1 - n^{-10}$  the order of the giant is close to its expectation. Here we also bound the probable number of isolated vertices. Moreover, in Section 3.7.4 we compute  $\text{Var}(\mathcal{N}(H_d(n, p)))$ , and in Section 3.7.5 we prove (3.6). Finally, in Section 3.7.6 we carry out a technical computation needed in Section 3.7.1. The material in this section builds on the work on random graphs [30, Ch. 10] and [39, Ch. 5].

### 3.7.1 Preliminaries on Branching Processes

During the branching process there are two kinds of “organisms”: living and dead ones. In the beginning, there is only one organism, which is alive. In the  $i$ 'th epoch, a living organism is chosen, produces a number  $Z_i$  of children, and dies. Here the  $Z_i$ 's are mutually independent random variables. We say that the process *dies out* if at some epoch  $i$  the last living organism dies without producing any offspring. We will mainly be interested in the probability that the process dies out, and in the total number of organisms generated throughout the process before its extinction.

We shall consider three types of branching processes. In the first process  $\mathcal{P}_d(c)$ , the mutually independent random variables  $Z_i$  are distributed as multiples  $(d-1) \cdot \text{Po}(c)$  of a Poisson variable. Thus, the process can be described formally as follows: we let  $Y_0 = 1$ , and  $Y_i = Y_{i-1} + Z_i - 1$  for  $i \geq 1$ . Hence,  $Y_i$  is the number of living organisms at epoch  $i$ . Additionally, we let  $T$  be the least integer  $i \geq 1$  such that  $Y_i = 0$  if such a number  $i$  exists, and  $T = \infty$  otherwise. Then  $T$  equals the total number of organisms that occur before the process dies out.

In addition to  $\mathcal{P}_d(c)$ , we will deal with two further types of branching processes. In the process  $\mathcal{P}_d(n, p)$ , the number  $Z'_i$  of offspring generated at epoch  $i$  is distributed as a multiple  $(d-1)\text{Bin}(\binom{n-1}{d-1}, p)$  of a binomially distributed random variable; as before, the  $Z'_i$ 's are mutually

independent. Moreover, the number of living organisms at epoch  $i$  is  $Y'_0 = 1$ , and  $Y'_i = Y'_{i-1} + Z'_i - 1$  for  $i \geq 1$ . We let  $T'$  be the least  $i$  such that  $Y'_i = 0$  if such an  $i$  exists, and  $T' = \infty$  otherwise.

To define the third type of branching process, let  $0 \leq \alpha_0, \alpha_1, \alpha_2 \leq 1$  be such that  $\alpha_0 + \alpha_1 + \alpha_2 = 1$ , and  $\alpha_1 = \tilde{O}(n^{-1})$ ,  $\alpha_2 = \tilde{O}(n^{-2})$ . Let  $1 \leq l \leq n$ . We define a probability distribution  $\mathcal{B}_d(n, l, p)$  as follows:

$$\begin{aligned} \mathbb{P}[\mathcal{B}_d(n, l, p) = 0] &= \alpha_0 \cdot \mathbb{P}\left[\text{Bin}\left[\binom{n-l}{d-1}, p\right] = 0\right] + \alpha_1 \cdot \mathbb{P}\left[\text{Bin}\left[\binom{n-l}{d-1}, p\right] \leq 1\right] + \alpha_2, \quad (3.61) \\ \mathbb{P}[\mathcal{B}_d(n, l, p) = j] &= \alpha_0 \cdot \mathbb{P}\left[\text{Bin}\left[\binom{n-l}{d-1}, p\right] = j\right] + \alpha_1 \cdot \mathbb{P}\left[\text{Bin}\left[\binom{n-l}{d-1}, p\right] = l+1\right] \quad (j \geq 1). \end{aligned}$$

In other words,  $\mathcal{B}_d(n, p)$  is  $\text{Bin}\left(\binom{n-\tilde{O}(1)}{d-1}, p\right)$  with probability  $\alpha_0$ ,  $\mathcal{B}_d(n, p)$  is  $\max\{\text{Bin}\left(\binom{n-\tilde{O}(1)}{d-1} - 1, 0\right)$  with probability  $\alpha_1$ , and  $\mathcal{B}_d(n, p)$  equals the constant 0 with probability  $\alpha_2$ . Let  $(Z''_1, Z''_2, \dots)$  be a sequence of mutually independent random variables with distribution  $\mathcal{B}_d(n, p)$ . Then we define the branching process  $\mathcal{P}_d(n, l, p)$  similarly as before: we let  $Y''_0 = 1$ ,  $Y''_i = Y''_{i-1} + Z''_i - 1$  ( $i \geq 1$ ), and we let  $T''$  be the number of organisms generated before extinction.

We shall compare  $\mathcal{P}_d(n, p)$  and  $\mathcal{P}_d(n, l, p)$  for  $p = c/\binom{n-1}{d-1}$  with  $\mathcal{P}_d(c)$ . To this end, we first derive the relevant results for  $\mathcal{P}_d(c)$ . Then, we show that these results carry over to the other two processes.

In the sequel, we assume that  $(d-1)^{-1} + \Omega(1) < c = o(\ln n)$ . To study  $\mathcal{P}_d(c)$ , let  $q_i = \mathbb{P}[T = i]$ , and set  $q(X) = \sum_{1 \leq i < \infty} q_i X^i$ . In addition, let

$$r(X) = \sum_{0 \leq i < \infty} \mathbb{P}[(d-1) \cdot \text{Po}(c) = i] X^i = \sum_{0 \leq i < \infty} \frac{c^i}{i!} \exp(-c) X^{(d-1)i} = \exp[c(X^{d-1} - 1)]$$

be the generating function of the successor distribution  $(d-1) \cdot \text{Po}(c)$ . If in the process  $\mathcal{P}_d(c)$  the first organism has  $i$  successors, then the number total of organisms is one (the first organism) plus the total number of organisms in  $i$  mutually independent processes  $\mathcal{P}_d(c)$  (the processes initiated

by the  $i$  successors of the first organism). Therefore,  $q(X)$  satisfies the equation

$$q(X) = \sum_{0 \leq i < \infty} P[(d-1) \cdot \text{Po}(c) = i] X q(X)^i = X \cdot r(q(X)) = X \cdot \exp [c(q(X)^{d-1} - 1)]. \quad (3.62)$$

**Proposition 3** *Let  $1 - a = P[T < \infty]$  be the probability that the branching process dies out. Then  $1 - a$  is the unique solution of the equation  $X = \exp [c(X^{d-1} - 1)]$  that lies strictly between 0 and 1.*

**Proof** As  $1 - a = q(1)$ , the fact that  $1 - a$  satisfies  $1 - a = \exp [c((1 - a)^{d-1} - 1)]$  follows immediately from (3.62). Furthermore, the general theory of branching processes shows that  $0 < 1 - a < 1$ , because  $E[(d-1) \cdot \text{Po}(c)] = (d-1)c > 1$ , and that the equation  $X = \exp [c(X^{d-1} - 1)]$  has only one solution in this range (cf. [36, p. 297]).  $\square$

Let  $\mu = \sum_{1 \leq i < \infty} i q_i$ . Then  $(1 - a)^{-1} \mu$  is the expected number of organisms that occur during the process given that the process dies out.

**Proposition 4** *We have  $\mu = \sum_{1 \leq i < \infty} i q_i = (1 - a)(1 - c(d-1)(1 - a)^{d-1})^{-1}$ .*

**Proof** Consider the derivative  $q'(X) = \sum_{1 \leq i < \infty} i q_i X^{i-1}$ . Differentiating both sides of (3.62), we get  $q'(X) = q(X) + (d-1)cq(X)^{d-1}q'(X)$ . Hence,  $q'(X) = q(X)(1 - c(d-1)q(X)^{d-1})^{-1}$ . As  $\mu = q'(1)$  and  $1 - a = q(1)$ , we obtain the desired formula.  $\square$

**Lemma 12** *There is a number  $\beta_c = \Omega_c(1)$  such that for all sufficiently large  $N$  we have  $P[N < T < \infty] \leq \exp(-\beta_c N)$ .*

**Proof** We can bound  $P[N < T < \infty]$  as follows: if  $N < T < \infty$ , then either the number  $\sum_{i=1}^N Z_i - N$  of living organisms at epoch  $N + 1$  is less than  $\frac{1}{2}N [(d-1)c - 1]$ , or there are

at least  $\frac{1}{2}N[(d-1)c-1]$  living organisms, but all the branching processes initiated by these organisms will die out. Hence,

$$\mathbb{P}[N < T < \infty] \leq \mathbb{P}\left[\sum_{i=1}^N Z_i \leq \frac{N[(d-1)c+1]}{2}\right] + (1-a)^{N[(d-1)c-1]/2}. \quad (3.63)$$

Furthermore,  $(d-1)^{-1}\sum_{i=1}^N Z_i$  has Poisson distribution with mean  $Nc$ . Therefore, letting  $\lambda = ((d-1)c-1)(2(d-1)c)^{-1}$  and applying Chernoff bounds, we obtain

$$\mathbb{P}\left[\sum_{i=1}^N Z_i \leq \frac{N[(d-1)c+1]}{2}\right] \leq \mathbb{P}\left[(d-1)^{-1}\sum_{i=1}^N Z_i \leq (1-\lambda)Nc\right] \leq \exp\left[-\frac{\lambda^2 c N}{2}\right]. \quad (3.64)$$

Combining (3.63) and (3.64) completes the proof.  $\square$

**Lemma 13** We have  $1-a = \exp(-\Theta(c))$ .

**Proof** Of course,  $\mathbb{P}[T < \infty] \geq \mathbb{P}[Z_1 = 0] = \exp(-c)$ . On the other hand, suppose that  $c \geq 3(d-1)^{-1}$ . Then the Chernoff bound (3.8) yields  $\mathbb{P}[T < \infty] \leq \mathbb{P}[Z_1 \leq \frac{c}{2}] + \mathbb{P}[T < \infty | Z_1 \geq \frac{c}{2}] \leq \exp(-\Omega(c)) + (1-a)^{c/2}$ . As  $1-a$  is bounded away from 1, we have  $(1-a)^{c/2} \leq \exp(-\Omega(c))$ .

$\square$

Now, we shall compare  $\mathcal{P}_d(c)$  with  $\mathcal{P}_d(n, p)$  and  $\mathcal{P}_d(n, l, p)$ , where  $p = c/\binom{n-1}{d-1}$  and  $l = \tilde{O}(1)$ .

**Proposition 5** Let  $\ln^2 n \leq N = \tilde{O}(1)$ . Set  $\mu' = \sum_{1 \leq i \leq N} i \cdot \mathbb{P}[T' = i]$ ,  $\mu'' = \sum_{1 \leq i \leq N} i \cdot \mathbb{P}[T'' = i]$ . Then  $\mu', \mu'' \sim \mu$  and  $\mathbb{P}[T' \leq N] = \mathbb{P}[T < \infty] + \tilde{O}(n^{-1})$ ,  $\mathbb{P}[T'' \leq N] = \mathbb{P}[T < \infty] + \tilde{O}(n^{-1})$ .

To prove Proposition 5, we employ the following lemma.

**Lemma 14** *Suppose that  $k = \tilde{O}(1)$ . Then*

$$P(T' = k) = (1 + \tilde{O}(n^{-1})) \cdot P(T = k), \quad (3.65)$$

$$P(T'' = k) = (1 + \tilde{O}(n^{-1})) \cdot P(T = k). \quad (3.66)$$

The proof of Lemma 14 is just a tedious computation, which we defer to Section 3.7.6.

*Proof of Proposition 5.* Lemma 14 immediately yields that  $P[T' \leq N] = (1 + \tilde{O}(n^{-1}))P[T \leq N]$ . Further, by Lemma 12 we have  $P[N < T < \infty] = \tilde{O}(n^{-1})$ . Hence,  $P[T' \leq N] = (1 + \tilde{O}(n^{-1}))P[T < \infty]$ . To prove that  $\mu' \sim \mu$ , note that  $\mu' = (1 + \tilde{O}(n^{-1})) \sum_{1 \leq i \leq N} i \cdot P[T = i]$ . Moreover, by Lemma 12 we have

$$\mu - \sum_{1 \leq i \leq N} i \cdot P[T = i] = \sum_{N < i < \infty} i \cdot P[T = i] \leq \sum_{N < i < \infty} i \cdot \exp(-\beta_c i) = \exp(-\Omega_c(N)),$$

so that  $\mu \sim \mu'$ . Similar arguments applies to  $T''$  and  $\mu''$ .

### 3.7.2 Exploring the Components of $H_d(n, p)$

Throughout, we set  $p = c/\binom{n-1}{d-1}$ ,  $m = cn/d$ , and we suppose that  $c = o(\ln n)$ . We use the branching processes from Section 3.7.1 to investigate the order of the largest component of  $H_d(n, p)$ . More precisely, we shall employ the branching processes to approximate the following *search process*, which explores the connected component of a vertex  $v$  in  $H = H_d(n, p)$ . During the search process, the vertices of  $H$  are either dead, alive, or neutral. Initially, only  $v$  is alive, and all other vertices are neutral. In each step, a living vertex  $w$  is chosen arbitrarily. We investigate all edges  $e$  of  $H$  that contain  $w$ , at least one neutral vertex, but no dead vertex (since a dead vertex indicates that the edge has already been explored). All neutral vertices contained in such edges are made live, and  $w$  dies. When there are no living vertices left, the set of dead vertices is precisely the component of  $v$ .

Chapter 3. Counting Connected Graphs and Hypergraphs via the Probabilistic Method

Let  $Z_i^*$  signify the number of vertices made live in the  $i$ 'th epoch of the search process, and let  $Y_i^*$  be the number of living vertices at epoch  $i$ . Moreover, let  $T^* = \min\{i \geq 1 : Y_i^* = 0\}$  be the order of the component of  $v$ . The following proposition relates the search process with the branching processes from the previous section.

**Proposition 6** *For all epochs  $i$  of the search process  $Z_i^*$  is dominated by  $Z_i' = (d-1)\text{Bin}(\binom{n-1}{d-1}, p)$ . Furthermore, if  $n - l$  is the number of neutral vertices at epoch  $i$ , then  $Z_i^*$  dominates  $Z_i'' = \mathcal{B}_d(n, l, p)$  (where  $\mathcal{B}_d(n, l, p)$  is defined in (3.61)).*

We call a vertex  $v$   $x$ -cyclic if  $v$  has at most  $(d-1)d_H(v) - x$  neighbors (recall that the degree  $d_H(v)$  of  $v$  is the number of edges that contain  $v$ ). To prove Proposition 6, we need the following observation.

**Lemma 15** 1. *With probability  $\geq 1 - n^{-10}$  the maximal degree of  $H_d(n, p)$  is  $O(\ln n)$ .*

2. *The probability that a fixed vertex is  $x$ -cyclic is  $\leq \tilde{O}(n^{-x})$  ( $x = 1, 2$ ).*

**Proof** As the expected average degree is  $c = o(\ln n)$ , the first claim follows immediately from the Chernoff bound (3.8). Furthermore, the probability that a vertex  $v$  of degree  $g = O(\ln n)$  is  $x$ -cyclic is  $\leq \binom{g(d-1)}{x+1} n^{-x} = \tilde{O}(n^{-x})$ .  $\square$

*Proof of Proposition 6.* Let  $w$  be the vertex whose neighbors are explored at the  $i$ 'th epoch. Since the number of edges of  $H_d(n, p)$  that contain  $w$  is distributed as  $\text{Bin}(\binom{n-1}{d-1}, p)$ , the total number of neighbors of  $w$  is dominated by  $Z_i' = (d-1)\text{Bin}(\binom{n-1}{d-1}, p)$ . Consequently,  $Z_i'$  also dominates the number  $Z_i^*$  of neighbors of  $w$  that are neutral at epoch  $i$ .

Further, let  $\alpha_x$  be the probability that  $w$  is  $x$ -cyclic ( $x = 1, 2$ ), and let  $\alpha_0 = 1 - \alpha_1 - \alpha_2$ . The number of edges that contain  $w$  and  $d-1$  neutral vertices is distributed as  $\text{Bin}(\binom{n-1}{d-1}, p)$ . Moreover, if  $w$  is not cyclic, then every such edge contributes  $d-1$  to  $Z_i^*$ , so that  $Z_i^*$  dominates  $(d-1)\text{Bin}(\binom{n-1}{d-1}, p)$ . Moreover, if  $w$  is 1-cyclic but not 2-cyclic, then  $Z_i^*$  dominates  $\max\{(d-$

$1)\text{Bin}\left(\binom{n-k}{d-1}, p\right) - 1, 0\}$ . Finally, if  $w$  is 2-cyclic, then at least  $Z_i^*$  dominates the constant random variable 0. Thus, by the estimates on  $\alpha_1$  and  $\alpha_2$  given in Lemma 15 and the definition (3.61) of  $\mathcal{B}_d(n, l, p)$ ,  $Z_i^*$  dominates  $Z_i'' = \mathcal{B}_d(n, l, p)$ .

By Proposition 6, the number  $Y_i^*$  of living vertices at the  $i$ 'th epoch of the search process is inbetween the numbers  $Y_i''$  and  $Y_i'$  of living organisms in the processes  $\mathcal{P}_d(n, p)$  and  $\mathcal{P}_d(n, l, p)$ . Hence,  $T^*$  dominates the random variable  $T''$ , and  $T^*$  is dominated by  $T'$ .

**Lemma 16** *If  $c < (d-1)^{-1} - \Omega(1)$ , then  $\mathcal{N}(H_d(n, p)) \leq k = \frac{3(d-1)^2}{(1-(d-1)c)^2} \ln n$  w.h.p.*

**Proof** Let  $v$  be a starting vertex. We explore the component of  $v$  via the search process. If the component of  $v$  has order  $\geq k$ , then  $\sum_{i=1}^k Z_i^* \geq k-1$ . Let  $(Z_i')_{i \geq 1}$  be a family of mutually independent random variables with distribution  $(d-1)\text{Bin}\left(\binom{n-1}{d-1}, p\right)$ . Then by Proposition 6 we have

$$\mathbb{P}\left[\sum_{i=1}^k Z_i^* \geq k-1\right] \leq \mathbb{P}\left[\sum_{i=1}^k Z_i' \geq k-1\right] = \mathbb{P}\left[\text{Bin}\left[k\binom{n-1}{d-1}, p\right] \geq \frac{k-1}{d-1}\right]. \quad (3.67)$$

The mean of the above binomial distribution is  $\lambda = k\binom{n-1}{d-1}p = kc$ , so that  $t = \frac{k-1}{d-1} - \lambda = \frac{k(1-(d-1)c)-1}{d-1}$ . Hence, the Chernoff bound (3.8) and the choice of  $k$  entail that

$$\mathbb{P}\left[\text{Bin}\left[k\binom{n-1}{d-1}, p\right] \geq \frac{k-1}{d-1}\right] \leq \exp\left[-\frac{t^2}{2(\lambda+t/3)}\right] \leq \exp\left[-\frac{9 \ln n}{2+o(1)}\right] \leq n^{-4}. \quad (3.68)$$

Combining (3.67) and (3.68), we conclude that for each vertex  $v$  the probability that  $v$  lies in a component of order  $> k$  is  $\leq n^{-4}$ . Thus, by the union bound all vertices lie in components of order  $\leq k$  w.h.p.  $\square$

Let us now consider the case  $c > (d-1)^{-1} + \Omega(1)$ . Let  $a, b$  be as in Theorem 5. The following lemma states that if we start the search process from a vertex  $v$ , then either the process will die out after inspecting only  $k_- = O_c(\ln n)$  vertices, or for all ‘‘intermediate’’ epochs  $k_- \leq k \leq n^{2/3}$  there will be plenty of living vertices.

**Lemma 17** Suppose that  $c > (d-1)^{-1} + \Omega(1)$ . Let  $k_- = \frac{1000(d-1)^2 c}{(1-(d-1)c)^2} \ln n$  and  $k_+ = n^{2/3}$ . Then

$$\mathbb{P} \left[ \forall k_- \leq k \leq k_+ : Y_k^* \geq \frac{1}{2}((d-1)c-1)k \mid T^* \geq k_- \right] \geq 1 - n^{-20}.$$

**Proof** Consider the random variable  $T^{**} = \min\{k \geq k_- : Y_k^* < \frac{1}{2}((d-1)c-1)k\}$ . Then  $T^{**} \leq \max\{T^*, k_-\}$ , and the assertion can be restated as

$$\mathbb{P} [T^{**} \leq k_+ \mid T^* \geq k_-] \leq n^{-20}. \quad (3.69)$$

Now, for all  $k_- \leq k \leq k_+$  we have

$$\mathbb{P} [T^{**} = k \mid T^* \geq k_-] \leq \mathbb{P} \left[ \sum_{i=1}^k Z_i^* < \frac{((d-1)c+1)k}{2} \mid T^* \geq k \right]. \quad (3.70)$$

Let  $(Z''_i)_{i \geq 1}$  be a family of mutually independent  $\mathcal{B}_d(n, k_+, p)$ -distributed random variables. Then by Proposition 6,  $Z_i^*$  dominates  $Z''_i$  for all  $1 \leq i \leq k$ , so that (3.70) yields

$$\mathbb{P} [T^{**} = k \mid T^* \geq k_-] \leq \mathbb{P} \left[ \sum_{i=1}^k Z''_i < \frac{((d-1)c+1)k}{2} \right]. \quad (3.71)$$

By the definition (3.61) of the distribution  $\mathcal{B}_d(n, k_+, p)$ , each  $Z''_i$  is distributed as  $(d-1)\text{Bin}(\binom{n-k_+}{d-1}, p)$  with probability  $\alpha_0 = 1 - \tilde{O}(n^{-1})$ , and with probability  $1 - \alpha_0$   $Z''_i$  has some other distribution but dominates the constant random variable 0. Hence,

$$\mathbb{P} \left[ \text{there are } \geq 100 \text{ indices } i \text{ such that } Z''_i \neq (d-1)\text{Bin}\left(\binom{n-k_+}{d-1}, p\right) \right] \leq \binom{k_+}{100} \alpha_0^{100} \leq n^{-30}. \quad (3.72)$$

Moreover, if  $Z''_i$  is distributed as  $(d-1)\text{Bin}(\binom{n-k_+}{d-1}, p)$  for at least  $k-100$  indices  $i$ , then  $\sum_{i=1}^k Z''_i$  dominates  $(d-1)\text{Bin}(\binom{n-k_+}{d-1}, p)$ . Hence, combining (3.71) and (3.72), we get

$$\mathbb{P} [T^{**} = k \mid T^* \geq k_-] \leq \mathbb{P} \left[ \text{Bin} \left[ (k-100) \binom{n-k_+}{d-1}, p \right] < \frac{((d-1)c+1)k}{2(d-1)} \right] + n^{-30}. \quad (3.73)$$

Further, the mean of the above binomial distribution satisfies  $\lambda = (k - 100) \binom{n - k_+}{d-1} p = c(k + o(1))$ , so that  $t = \lambda - \frac{((d-1)c+1)k}{2(d-1)} = \frac{(d-1)c-1-o(1)}{2(d-1)} \cdot k$ . Therefore, the Chernoff bound (3.8) entails that

$$\begin{aligned} \mathbb{P} \left[ \text{Bin} \left[ (k - 100) \binom{n - k_+}{d-1}, p \right] < \frac{((d-1)c+1)k}{2(d-1)} \right] &\leq \exp \left[ -\frac{t^2}{2\lambda} \right] \\ &\leq \exp \left[ -\frac{[(d-1)c-1-o(1)]^2 k}{(8+o(1))(d-1)^2 c} \right]. \end{aligned} \quad (3.74)$$

The term on the right hand side of (3.74) is maximized for  $k = k_-$ , so that our choice of  $k_-$  ensures that

$$\frac{((d-1)c-1-o(1))^2 k}{(8+o(1))(d-1)^2 c} \geq 30 \ln n. \quad (3.75)$$

Combining (3.73)–(3.75), we get  $\mathbb{P} [T^{**} = k \mid T^* \geq k_-] \leq 2n^{-30}$ . Thus, (3.69) follows from the union bound:  $\mathbb{P} [T^{**} \leq k_+ \mid T^* \geq k_-] \leq \sum_{k=k_-}^{k_+} \mathbb{P} [T^{**} = k \mid T^* \geq k_-] \leq n^{2/3} \cdot 2n^{-30} \leq n^{-20}$ .

□

**Corollary 1** *If  $c > (d-1)^{-1} + \Omega(1)$ , then with probability  $\geq 1 - n^{-19}$  the random hypergraph  $H_d(n, p)$  has no component of order  $k_- \leq k \leq k_+$ .*

**Proof** Start the search process at a vertex  $v$ . If the process explores at least  $k_-$  vertices, then Lemma 17 entails that with probability  $\geq 1 - n^{-20}$  the process will indeed explore more than  $k_+$  vertices. Hence, with probability  $\geq 1 - n^{-19}$  there is no vertex  $v$  that belongs to a component of order  $k_- \leq k \leq k_+$ . □

**Corollary 2** *If  $c > (d-1)^{-1} + \Omega(1)$ , then with probability  $\geq 1 - n^{-16}$  in  $H_d(n, p)$  there is at most one component of order  $\geq n^{2/3}$ .*

**Proof** Start the search process at two vertices  $v', v''$ , and assume that the process runs for at least  $k_+$  epochs in both cases. Let  $V', V''$  be the encountered vertex sets up to epoch  $k_+$ . If

$V' \cap V'' \neq \emptyset$ , then the components of  $v$  and  $v'$  coincide. Otherwise, consider the sets  $U' \subset V'$  and  $U'' \subset V''$  of living vertices. Let  $u' = |U'|$  and  $u'' = |U''|$ . Then Lemma 17 entails that with probability  $\geq 1 - n^{-19}$  we have  $u', u'' \geq \Omega_c(n^{2/3})$ . Moreover, let  $\nu$  be the number of all possible edges that contain vertices in both  $U'$  and  $U''$ . Then the probability that in  $H_d(n, p)$  there is no edge joining  $U'$  and  $U''$  is  $(1 - p)^\nu$ . If  $u', u'' = \Omega_c(n^{2/3})$ , then  $\nu = \Omega_c(n^{d-2/3})$ , so that  $(1 - p)^\nu \leq \exp[-\Omega_c(cn^{1/3})] < n^{-20}$ . Hence, the probability that  $v', v''$  belong to different components of order  $\geq n^{2/3}$  is  $\leq n^{-18}$ . Consequently, applying the union bound to all possible pairs of start vertices  $(v, v')$ , we conclude that with probability  $\geq 1 - n^{-16}$  there is at most one component of order  $\geq n^{2/3}$ .  $\square$

We call a vertex *big* if it lies in a component of order  $\geq n^{2/3}$ , and *small* if it lies in a component with  $\leq \ln^4 n$  vertices. Let  $a = a(c)$  be the probability that the branching process  $\mathcal{P}_d(c)$  dies out (cf. Proposition 3).

**Corollary 3** *Suppose that  $c > (d - 1)^{-1} + \Omega(1)$ . Then the expected number of small vertices is  $(1 - a)n + \tilde{O}(1)$ .*

**Proof** Let  $r$  be the probability that a fixed vertex  $v$  is small. Then by Proposition 6,  $r$  is bounded from above by the probability  $\rho''$  that the branching process with successor distribution  $\mathcal{B}_d(n, p, \ln^4 n)$  dies out. On the other hand, once more by Proposition 6,  $r$  is bounded from below by the probability  $\rho'$  that the branching process with distribution  $(d - 1)\text{Bin}(\binom{n-1}{d-1}, p)$  dies after at most  $N = \ln^4 n$  steps. By Proposition 5,  $|(1 - a) - \rho'|, |(1 - a) - \rho''| \leq \tilde{O}(1/n)$ , whence  $|r - (1 - a)| \leq \tilde{O}(1/n)$ .  $\square$

Finally, we determine the expected number of edges outside the giant of  $H_d(n, p)$ . Let  $m = \binom{n}{d}p$ .

**Lemma 18** *Suppose that  $c > (d - 1)^{-1} + \Omega(1)$ . Then the expected number of edges outside the largest component is  $(1 - a)^d m + n^{o(1)}$ .*

**Proof** To determine the expected number of edges outside the giant, we may assume that  $H = H_d(n, p)$  is such that all vertex degrees are  $O(\ln n)$  (by Lemma 15), and that the largest component of  $H$  is the unique component that has order  $> N = \ln^4 n$  (by Corollaries 1 and 2). Our goal is to estimate the probability  $\rho_g$  that the search process starting from a vertex  $v$  dies after  $\leq N$  steps, given that the degree of  $v$  is  $g = O(\ln n)$ . Let  $v_1, \dots, v_g$  be the neighbors of  $v$ .

On the one hand, Proposition 6 entails that the probability  $\sigma_g$  that  $g$  independent branching processes  $\mathcal{P}_d(n, p)$  die after generating at most  $N$  organisms in total provides a lower bound on  $\rho_g$ . Indeed, let  $\sigma$  be the probability that  $\mathcal{P}_d(n, p)$  dies after generating at most  $N/g$  organisms. Then  $\sigma_g \geq \sigma^g$ . Hence, Proposition 5 entails that

$$\rho_g \geq \sigma_g \geq \sigma^g \geq (1 - a)^g - \tilde{O}(n^{-1}). \quad (3.76)$$

On the other hand, once more by Proposition 6, the probability  $\tau_g$  that  $g$  independent runs of a  $\mathcal{P}_d(n, N, p)$  branching processes die after producing  $\leq N$  organisms in total yields an upper bound on  $\rho_g$ . Let  $\tau$  signify the probability that one  $\mathcal{P}_d(n, N, p)$  dies out after producing  $\leq N$  organisms. Then  $\tau_g \leq \tau^g$ , so that Proposition 5 implies that

$$\rho_g \leq \tau_g \leq \tau^g \leq (1 - a)^g + \tilde{O}(n^{-1}). \quad (3.77)$$

Now, suppose that  $g = (d - 1)j$  is a multiple of  $d - 1$ . Then by Lemma 15

$$\text{P}[v \text{ has } g \text{ neighbors}] = \binom{n-1}{d-1} p^g (1-p)^{\binom{n-1}{d-1} - g} + \tilde{O}(n^{-1}) = \frac{c^j}{j!} \exp(-c) + \tilde{O}(n^{-1}). \quad (3.78)$$

Combining (3.76)–(3.78), we get  $\text{P}[v \text{ is small and has } g \text{ neighbors}] = (1 - a)^g \frac{c^j}{j!} \exp(-c) + \tilde{O}(n^{-1})$ . Further, since the expected number of edges outside the largest component is  $(1 - a)n + \tilde{O}(1)$  and  $1 - a = \exp(-\Theta(c)) = n^{-o(1)}$  by Lemma 13 and our assumption  $c = o(\ln n)$ , Lemma 15 implies that

$$\begin{aligned} \text{P}[v \text{ has degree } j | v \text{ is small}] &= \frac{\text{P}[v \text{ is small and has } g \text{ neighbors}] + \tilde{O}(n^{-1})}{\text{P}[v \text{ is small}]} \\ &= (1 - a)^{g-1} \frac{c}{j!} \exp(-c) + n^{o(1)-1}. \end{aligned}$$

Consequently, the expected average degree outside the giant is

$$\begin{aligned} \sum_{j=0}^{\ln^4 n} j \mathbb{P}[v \text{ has degree } j | v \text{ is small}] &= n^{o(1)-1} + \sum_{j=0}^{\ln^4 n} j(1-a)^{(d-1)j-1} \frac{c^j}{j!} \exp(-c) \\ &= n^{o(1)-1} + (1-a)^{d-2} c \sum_{j=1}^{\infty} (1-a)^{(d-1)(j-1)} \frac{c^{j-1}}{(j-1)!} \exp(-c) = (1-a)^{d-1} c + n^{o(1)-1}, \end{aligned}$$

where the last step follows from the equation  $1-a = \exp(c((1-a)^{d-1} - 1))$  that defines  $a$ .  $\square$

In summary, Lemma 16 establishes the first part of Theorem 5, Lemma 13 shows (3.3), (3.4) follows from Corollary 3 and Lemma 18. Finally, the assertion concerning the second largest component of  $H_d(n, p)$  follows from Corollaries 1 and 2.

### 3.7.3 Large Deviations of $\mathcal{N}(H_d(n, p))$ and the Number of Isolated Vertices

We assume that  $(d-1)^{-1} + \Omega(1) < c = o(\ln n)$  and let  $a$  be as in Theorem 5. We set  $p = c / \binom{n-1}{d-1}$ . Let  $N = \ln^4 n$ . In this section we prove the following.

**Proposition 7** *With probability  $\geq 1 - \exp(-n^{\Omega(1)})$  the random hypergraph  $H_d(n, p)$  has  $n \cdot \exp(-c) + o(n^{9/10})$  isolated vertices. Moreover, with probability  $\geq 1 - n^{-10}$  the number of vertices in components of order  $\leq N$  is  $(1 + o(1))(1-a)n$ .*

To prove the lemma, we replace  $H_d(n, p)$  by the following closely related probability space. Let  $M = n^{1.1}$ , and let  $T = (e_1, \dots, e_M)$  be a random  $M$ -tuple of edges chosen uniformly at random among all  $\binom{n}{d}^M$  possible tuples. Let  $X = |\{e_1, \dots, e_M\}|$  be the number of distinct edges occurring in  $T$ . Moreover, let  $Y = \text{Bin}(\binom{n}{d}, p)$  be independent of  $T$ . Then, we define a hypergraph  $H(T, Y)$  with vertex set  $V$  as follows:  $H(T, Y)$  consists of the first  $Y$  distinct edges of the tuple  $T$  if  $Y \leq X$ , and  $H(T, Y)$  consist of all edges  $\{e_1, \dots, e_M\}$  otherwise. Thus,  $H(T, Y)$  has  $\min\{X, Y\}$  edges. Furthermore, given its number of edges,  $H(T, Y)$  is

uniformly distributed. As the number of edges of  $H_d(n, p)$  is  $\text{Bin}\left(\binom{n}{d}, p\right)$ , we have the following observation:

the conditional distribution of  $H(T, Y)$  given that  $Y \leq \frac{M}{2} \leq X$  coincides with the conditional distribution of  $H_d(n, p)$  given that the number of edges is  $\leq \frac{M}{2}$ . (3.79)

The benefit of the  $H(T, Y)$  model is that we can easily prove the following concentration result.

**Lemma 19** *Suppose that  $R$  is a function from the set of all hypergraphs with vertex set  $V$  to the interval  $[0, n^d]$  that satisfies the following Lipschitz condition: if a hypergraph  $H'$  is obtained from  $H$  by adding or deleting one edge, then  $|R(H) - R(H')| \leq \sigma = \tilde{O}(1)$ . Then*

$$\mathbb{P} [ |R(H(T, Y)) - \mathbb{E}[R(H(T, Y))] | \geq n^{3/4} ] \leq \exp(-n^{\Omega(1)}).$$

**Proof** As  $Y = \text{Bin}\left(\binom{n}{d}, p\right)$ , the Chernoff bound (3.8) yields that

$$\mathbb{P} [ |Y - \mathbb{E}(Y)| > n^{2/3} ] \leq \exp(-n^{\Omega(1)}). \quad (3.80)$$

Moreover, the Lipschitz condition ensures that for all numbers  $y$  such that  $|y - \mathbb{E}(Y)| \leq n^{2/3}$  we have

$$| \mathbb{E}[R(H(T, y))] - \mathbb{E}[R(H(T, Y))] | \leq (1 + o(1))n^{2/3}\sigma; \quad (3.81)$$

for by (3.80) with probability  $1 - \exp(-n^{\Omega(1)})$  the hypergraph  $H(T, Y)$  can be obtained from  $H(T, y)$  by adding or deleting at most  $n^{2/3}$  edges, so that  $|R(H(T, Y)) - R(H(T, y))| \leq \sigma n^{2/3}$ .

Furthermore, since the components of the random  $M$ -tuple  $T$  are mutually independent, Azuma's inequality (cf. [39, Corollary 2.27]) entails that for all such  $y$  we have

$$\mathbb{P} [ |R(H(T, y)) - \mathbb{E}[R(H(T, y))] | \geq n^{2/3} ] \leq 2 \exp \left[ -\frac{n^{4/3}}{2M\sigma^2} \right] \leq \exp(-n^{\Omega(1)}). \quad (3.82)$$

Thus, the assertion follows from (3.80)–(3.82).  $\square$

The following lemma shows that the conditioning (3.79) is not very restrictive, i.e., that  $H(T, Y)$  is “essentially the same” as  $H_d(n, p)$ .

**Lemma 20** We have  $\mathbb{P}[|E(H_d(n, p))| \leq \frac{M}{2}] \geq 1 - \exp(-n)$  and  $\mathbb{P}[Y \leq \frac{M}{2} \leq X] \geq 1 - \exp(-n^{\Omega(1)})$ .

**Proof** As  $|E(H_d(n, p))| = \text{Bin}(\binom{n}{d}, p)$ , the first assertion follows immediately from  $c = o(\ln n)$  and the Chernoff bound (3.8), which also implies that  $\mathbb{P}[Y > \frac{M}{2}] = o(\exp(-n))$ . As  $E(X) \sim M$  and  $X$  satisfies the Lipschitz condition in Lemma 19 with  $\sigma = 1$ , Lemma 19 yields  $\mathbb{P}[X < \frac{M}{2}] = \exp(-n^{\Omega(1)})$ .  $\square$

*Proof of Proposition 7.* By Lemma 20 and (3.79), it suffices to prove that the desired estimates hold for  $H(T, Y)$ . Let  $I$  be the number of isolated vertices in  $H(T, Y)$ , and let  $Z$  be the number of vertices in components of order  $\leq N$ . Since the degree of each vertex in  $H_d(n, p)$  has distribution  $\text{Bin}(\binom{n-1}{d-1}, p)$ , the expected number of isolated vertices in  $H_d(n, p)$  is  $(1 + o(1))n \exp(-c)$ . Thus, due to (3.79) and Lemma 20 we have  $E(I) \sim n \exp(-c)$  as well. Similarly, Corollaries 1–3 entail that  $E(Z) \sim (1 - a)n$ . Now,  $I$  (resp.  $Z$ ) enjoys the Lipschitz condition in Lemma 19 with  $\sigma = 2$  (resp.  $\sigma = 2N = \tilde{O}(1)$ ), so that the assertion follows from Lemma 19.

Proposition 7 implies the assertions about the probability that there exists a component of order  $(1 + o(1))an$  and about the number of isolated vertices of  $H_d(n, p)$  in Theorem 5.

### 3.7.4 The Variance of $\mathcal{N}(H_d(n, p))$

Suppose that  $(d - 1)^{-1} + \Omega(1) < c = o(\ln n)$ , let  $p = c / \binom{n-1}{d-1}$ , and let  $a$  be as in Theorem 5. In this section, we use the analogy between the search process in  $H_d(n, p)$  and the branching processes established in Sections 3.7.1 and 3.7.2 to compute  $\text{Var}(\mathcal{N}(H_d(n, p)))$ . For a vertex  $v$  of  $H_d(n, p)$  we let  $C_v$  denote the vertex set of the connected component of  $v$ , and  $N_v = |C_v|$ . As in Section 3.7.2, we say that a vertex  $v$  is *small* if  $N_v \leq N = \ln^4 n$ . Let  $\rho = \mathbb{P}[v \text{ is small}]$  and  $\mu^* = \sum_{x=1}^N x \mathbb{P}[N_v = x]$ . Moreover, let  $0 < a < 1$  be as in Theorem 5. Then by Propositions 4, 5 and 6 we have

$$\mu^* \sim \mu, \quad \rho \sim 1 - a. \tag{3.83}$$

**Lemma 21** *Let  $v, w \in V$  be distinct vertices. Then*

$$\mathbb{P}[\text{both } v, w \text{ are small}] - \rho^2 \sim n^{-1} [\mu(1 + ((d-1)c - 1)\mu) - a(1-a)].$$

**Proof** Since  $\rho = \sum_{x=1}^N \mathbb{P}[N_v = x]$ , we have

$$\begin{aligned} \mathbb{P}[v, w \text{ are small}] - \rho^2 &= \left[ \sum_{x=1}^N \mathbb{P}[w \text{ is small} | N_v = x] \cdot \mathbb{P}[N_v = x] \right] - \rho^2 \\ &= \sum_{x=1}^N (\mathbb{P}[w \text{ is small}, C_v \cap C_w = \emptyset | N_v = x] + \mathbb{P}[w \in C_v | N_v = x] - \rho) \mathbb{P}[N_v = x]. \end{aligned}$$

We split the last sum in two terms, corresponding to the cases  $w \notin C_v$  and  $w \in C_v$ :

$$\begin{aligned} S_1 &= \sum_{x,y=1}^N (\mathbb{P}[N_w = y | w \notin C_v, N_v = x] - \mathbb{P}[N_w = y]) \mathbb{P}[w \notin C_v | N_v = x] \mathbb{P}[N_v = x], \\ S_2 &= (1 - \rho) \sum_{x=1}^N \mathbb{P}[w \in C_v | N_v = x] \mathbb{P}[N_v = x]. \end{aligned}$$

Now,  $\mathbb{P}[v, w \text{ are small}] - \rho^2 = S_1 + S_2$ , and we shall compute  $S_1$  and  $S_2$  separately.

To estimate  $S_2$ , note that  $\mathbb{P}[w \in C_v | N_v = x] = \binom{n-2}{x-2} \binom{n-1}{x-1}^{-1} = \frac{x-1}{n-1}$ ; for given that  $w \in C_v$ , there are  $\binom{n-2}{x-2}$  ways to choose the remaining  $x-2$  vertices in  $C_v$ , while the total number of ways to choose  $C_v$  is  $\binom{n-1}{x-1}$ . Hence,

$$S_2 \sim (1 - \rho)n^{-1} \sum_{x=1}^N (x-1) \mathbb{P}[N_v = x] = (1 - \rho)n^{-1} [\mu^* - \rho] \stackrel{(3.83)}{\sim} n^{-1} a(\mu - (1-a)). \quad (3.84)$$

In order to estimate  $S_1$ , we observe that

$$\mathbb{P}[N_w = y | N_v = x, w \notin C_v] = \mathbb{P}[N_w = y \text{ in } H_d(n, p) - C_v]. \quad (3.85)$$

Given that  $N_v = x$ ,  $H_d(n, p) - C_v$  is distributed as a random hypergraph  $H_d(n-x, p)$ . Hence, the probability that  $N_w = y$  in  $H_d(n, p) - C_v$  equals the probability that a given vertex of  $H_d(n-x, p)$

belongs to a component of size  $y$ . Therefore, we can compare  $\mathbb{P}[N_w = y \text{ in } H_d(n, p) - C_w]$  and  $\mathbb{P}[N_w = y \text{ in } H_d(n, p)]$  as follows: in  $H_d(n - x, p)$  there are  $\binom{n-x-1}{y-1}$  ways to choose the set  $C_w \setminus \{y\}$ . Moreover, there are  $\binom{n-x}{d} - \binom{n-x-y}{d} - \binom{y}{d}$  possible edges connecting the chosen set  $C_w$  with  $V \setminus C_w$ , and as  $C_w$  is a component, none of these edges is present. Since each such edge is present with probability  $p$  independently, the probability that there is no  $C_w$ - $V \setminus C_w$  edge equals

$$(1 - p)^{\binom{n-x}{d} - \binom{n-x-y}{d} - \binom{y}{d}}.$$

By comparison, in  $H_d(n, p)$  there are  $\binom{n-1}{y-1}$  ways to choose the vertex set of  $C_w$ . Further, there are  $\binom{n}{d} - \binom{n-y}{d} - \binom{y}{d}$  possible edges connecting  $C_w$  and  $V \setminus C_w$ , each of which is present with probability  $p$  independently. Thus, letting  $\gamma = \binom{n-x}{d} - \binom{n-x-y}{d} - [\binom{n}{d} - \binom{n-y}{d}]$ , we obtain

$$\frac{\mathbb{P}[N_w = y \text{ in } H_d(n, p) - C_w]}{\mathbb{P}[N_w = y \text{ in } H_d(n, p)]} = \binom{n-x-1}{y-1} \binom{n-1}{y-1}^{-1} (1-p)^\gamma. \quad (3.86)$$

Concerning the quotient of the binomial coefficients, we have

$$\binom{n-x-1}{y-1} \binom{n-1}{y-1}^{-1} = \exp \left[ -\frac{x(y-1)}{n} + \tilde{O}(n^{-2}) \right]. \quad (3.87)$$

Moreover,  $\gamma = \binom{n}{d} \left[ \frac{\binom{n-x}{d} + \binom{n-y}{d} - \binom{n-x-y}{d}}{\binom{n}{d}} - 1 \right]$ . Expanding the falling factorials, we get

$$\gamma = \binom{n}{d} \left[ \binom{d}{2} n^{-2} (x^2 + y^2 - (x+y)^2) + \tilde{O}(n^{-3}) \right] = -\binom{n}{d-2} xy + \tilde{O}(n^{d-3}). \quad (3.88)$$

Plugging (3.87) and (3.88) into (3.86), we obtain

$$\begin{aligned} \frac{\mathbb{P}[N_w = y \text{ in } H_d(n, p) - C_w]}{\mathbb{P}[N_w = y \text{ in } H_d(n, p)]} &= \exp \left[ -\frac{x(y-1)}{n} + \tilde{O}(n^{-2}) \right] (1-p)^{-\binom{n}{d-2} xy + \tilde{O}(n^{d-3})} \\ &\stackrel{(3.7)}{=} \exp \left[ -\frac{x(y-1)}{n} + \binom{n}{d-2} xyp + \tilde{O}(n^{-2}) \right] = 1 + n^{-1} [((d-1)c - 1)xy + x] + \tilde{O}(n^{-2}). \end{aligned}$$

Therefore, by (3.85)

$$\begin{aligned} \mathbb{P}[N_w = y | N_v = x, w \notin C_v] - \mathbb{P}[N_w = y \text{ in } H_d(n, p)] \\ = \mathbb{P}[N_w = y \text{ in } H_d(n, p)] \left[ n^{-1} [((d-1)c - 1)xy + x] + \tilde{O}(n^{-2}) \right]. \end{aligned} \quad (3.89)$$

Substituting (3.89) into  $S_1$ , we get

$$\begin{aligned}
 S_1 &= \sum_{x,y=1}^N \mathbb{P}[N_v = x] \mathbb{P}[N_w = y] \cdot \left[ n^{-1} [((d-1)c-1)xy + x] + \tilde{O}(n^{-2}) \right] \\
 &\stackrel{(3.83)}{\sim} n^{-1} [((d-1)c-1)\mu^2 + \mu(1-a)]. \tag{3.90}
 \end{aligned}$$

Combining (3.84) and (3.90) and recalling that  $\mathbb{P}[v, w \text{ are small}] - \rho^2 = S_1 + S_2$  yields the desired result.  $\square$

**Corollary 4** *The variance of the number of small vertices is  $(1 + o(1))n\mu(1 + ((d-1)c-1)\mu)$ .*

**Proof** Let  $S$  be the number of small vertices in  $H_d(n, p)$ . Then by Lemma 21 and (3.83), we have

$$\begin{aligned}
 \text{Var}(S) &= \mathbb{E}(S^2) - \mathbb{E}(S)^2 = \sum_{v,w \in V} \mathbb{P}[\text{both } v, w \text{ are small}] - \mathbb{P}[v \text{ is small}] \mathbb{P}[w \text{ is small}] \\
 &= \sum_{v,w \in V, v \neq w} (\mathbb{P}[\text{both } v, w \text{ are small}] - \rho^2) + \sum_{v \in V} (\mathbb{P}[v \text{ is small}] - \rho^2) \\
 &\sim n [\mu(1 + ((d-1)c-1)\mu) - a(1-a)] + n [(1-a) - (1-a)^2] = n\mu(1 + ((d-1)c-1)\mu),
 \end{aligned}$$

as claimed.  $\square$

By Corollaries 1 and 2 with probability  $\geq 1 - n^{-10}$  a vertex is either small or belongs to the unique giant component. Hence, plugging the expression for  $\mu$  from Proposition 4 into the formula in Corollary 4, we obtain the expression for  $\text{Var}(\mathcal{N}(H_d(n, p)))$  stated in Theorem 5.

### 3.7.5 The Variance of the Number of Edges Outside the Giant

We assume that  $(d-1)^{-1} + \Omega(1) < c = o(\ln n)$ , set  $p = c/\binom{n-1}{d-1}$ ,  $m = cn/d$ , and let  $a, b$  be as in Theorem 5. Recall from Section 3.7.2 that a vertex  $v$  of  $H_d(n, p)$  is *small* if  $v$  lies in a

component of order  $\leq N = \ln^4 n$ . We denote the vertex set of the component of  $v$  by  $C_v$ . In this section, by the *largest component* of a random hypergraph  $H$  we mean the lexicographically first component of  $H$  of order  $\mathcal{N}(H)$ . Moreover, we say that a set  $S$  of vertices of  $H$  is *little* if the following holds:  $S$  contains no vertex of the largest component of the hypergraph  $H'$  that is obtained by adding to  $H$  all possible edges  $e \subset S$ . We let  $C_S = \bigcup_{v \in S} C_v$ , and  $N_S$  denotes the number of vertices in  $C_S$ . If  $|S| = O(1)$ , then by Corollaries 1 and 2 we have

$$\mathbb{P}[N_S \leq N] = \mathbb{P}[S \text{ is little}] + O(n^{-10}). \quad (3.91)$$

Furthermore, let us observe that by definition the following holds:

$$\begin{aligned} &\text{given that } S \text{ is little in } H = H_d(n, p), \text{ each possible edge } e \subset S \text{ is present} \\ &\text{with probability } p \text{ independently.} \end{aligned} \quad (3.92)$$

In order to estimate the expected number of edges outside the largest component of  $H_d(n, p)$ , we need the following lemma.

**Lemma 22** *Let  $S \subset V$  be a set of cardinality  $s = O(1)$ . Then  $|\mathbb{P}[S \text{ is little}] - (1-a)^s| \leq n^{-\Omega(1)}$ .*

**Proof** Since  $\mathbb{E}(\mathcal{N}(H_d(n, p))) = an + n^{o(1)}$  and  $\text{Var}(\mathcal{N}(H_d(n, p))) = O_c(n^{1/2})$  by (3.4) and (3.5), Chebyshev's inequality entails that

$$\mathbb{P}[|\mathcal{N}(H_d(n, p)) - an| \leq n^{2/3}] \geq 1 - n^{-1/4}. \quad (3.93)$$

Moreover, by Corollaries 1 and 2 with probability  $\geq 1 - n^{-10}$   $H_d(n, p)$  has a unique component of order  $\Omega_c(n)$ . Given that this is the case and that  $y = \mathcal{N}(H_d(n, p))$  satisfies  $|y - an| \leq n^{2/3}$ , the vertex set  $Y$  of the largest component of  $H_d(n, p)$  is a uniformly distributed subset of  $V$  of cardinality  $y$ . Hence, the probability that  $S \cap Y = \emptyset$  equals

$$\binom{n-s}{y} \binom{n}{y} = (1 - y/n)^s + n^{-\Omega(1)} = (1-a)^s + n^{-\Omega(1)}. \quad (3.94)$$

Thus, the assertion follows from (3.93) and (3.94).  $\square$

**Lemma 23** Let  $e \subset V$ ,  $|e| = d$ . Then  $\eta = \sum_{x=d}^N x \cdot \mathbb{P}[N_e = x] \sim d\mu$ .

**Proof** Consider  $d$  branching processes  $(P'_i)_{1 \leq i \leq d}$  of type  $\mathcal{P}_d(n, p)$  and  $d$  processes  $(P''_i)_{1 \leq i \leq d}$  of type  $\mathcal{P}_d(n, N, p)$ ; these  $2d$  processes are mutually independent. Let  $T'_i$  (resp.  $T''_i$ ) signify the number of organisms produced in the process  $P'_i$  (resp.  $P''_i$ ). Then Proposition 6 entails that  $N_e$  dominates  $\sum_{i=1}^d T''_i$ , while  $N_e$  is dominated by  $\sum_{i=1}^d T'_i$ . Consequently,  $\sum_{x=d}^N x \cdot \mathbb{P}[\sum_{i=1}^d T''_i = x] \leq \eta \leq \sum_{x=d}^N x \cdot \mathbb{P}[\sum_{i=1}^d T'_i = x]$ . Thus, the assertion follows from Proposition 5.  $\square$

**Lemma 24** Let  $e, f \subset V$ ,  $|e| = |f| = d$ ,  $e \cap f = \emptyset$ . Then  $\mathbb{P}[\text{both } e, f \text{ are little}] - \theta^2 \sim \zeta$ , where  $\theta = \mathbb{P}[N_e \leq N]$  and  $\zeta = [a(1-a)^{d-1}d(\eta - d(1-a)^d) + (((d-1)c-1)\eta + (1-a)^d)\eta] \cdot n^{-1}$ .

**Proof** The proof is an adaptation of the proof of Lemma 21. By (3.91),

$$\begin{aligned} \mathbb{P}[\text{both } e, f \text{ are little}] - \theta^2 &\sim \mathbb{P}[f \text{ is little and } N_e \leq N] - \theta \cdot \mathbb{P}[N_e \leq N] \\ &= \sum_{x=d}^N (\mathbb{P}[f \text{ is little} | N_e = x] - \theta) \mathbb{P}[N_e = x] \\ &\stackrel{(3.91)}{\sim} \sum_{x=d}^N (\mathbb{P}[N_f \leq N, f \cap C_e = \emptyset | N_e = x] + \mathbb{P}[f \text{ is little, } f \cap C_e \neq \emptyset | N_e = x] - \theta) \mathbb{P}[N_e = x]. \end{aligned}$$

Thus, letting

$$\begin{aligned} S_1 &= \sum_{x=d}^N (\mathbb{P}[N_f \leq N | f \cap C_e = \emptyset, N_e = x] - \theta) \mathbb{P}[f \cap C_e = \emptyset | N_e = x] \mathbb{P}[N_e = x], \\ S_2 &= \sum_{x=d}^N (\mathbb{P}[f \text{ is little} | f \cap C_e \neq \emptyset, N_e = x] - \theta) \mathbb{P}[f \cap C_e \neq \emptyset | N_e = x] \mathbb{P}[N_e = x], \end{aligned}$$

we get  $\mathbb{P}[\text{both } e, f \text{ are little}] - \theta^2 \sim S_1 + S_2$ .

With respect to  $S_1$ , observe that for any  $d \leq x \leq N = o(n)$  we have

$$\mathbb{P}[f \cap C_e \neq \emptyset | N_e = x] = \sum_{j=1}^d \binom{d}{j} \binom{n-2d}{x-j-d} \binom{n-d}{x-d}^{-1}. \quad (3.95)$$

Chapter 3. Counting Connected Graphs and Hypergraphs via the Probabilistic Method

For if  $1 \leq |f| \cap C_e = j \leq d$ , then there are  $\binom{d}{j}$  ways to choose the  $j$  elements of  $f \cap C_e$ , and then  $\binom{n-2d}{x-j-d}$  ways to choose the  $x-j-d$  elements of  $C_e \setminus (e \cup f)$ . In particular, if  $d \leq x \leq N = o(n)$ , then  $P[f \cap C_e = \emptyset | N_e = x] \sim 1$ . Hence,

$$S_1 \sim \sum_{x,y=d}^N (P[N_f = y | C_f \cap C_e = \emptyset, N_e = x] - P[N_f = y])P[N_e = x]. \quad (3.96)$$

Furthermore, given that  $H = H_d(n, p)$  is such that  $N_e = x$  and  $C_f \cap C_e = \emptyset$ , the probability that  $N_f = y$  equals the probability that  $N_f = y$  in  $H - C_e$ , which is distributed as  $H_d(n-x, d)$ . Therefore, the same computation as in the proof of Lemma 21 (cf. Equation (3.89)) shows that

$$\begin{aligned} P[N_f = y | C_f \cap C_e = \emptyset, N_e = x] - P[N_f = y] \\ = P[N_f = y] \left[ n^{-1} [((d-1)c-1)xy + x] + \tilde{O}(n^{-2}) \right]. \end{aligned} \quad (3.97)$$

Plugging (3.97) into (3.96), we get

$$\begin{aligned} S_1 &\sim \sum_{x,y=d}^N P[N_e = x] P[N_f = y] \left[ n^{-1} [((d-1)c-1)xy + x] + \tilde{O}(n^{-2}) \right] \\ &\sim n^{-1} [((d-1)c-1)\eta + (1-a)^d] \eta, \end{aligned} \quad (3.98)$$

due to Lemma 23 and because  $\theta \sim (1-a)^d$  by Lemma 22 and (3.91).

Regarding  $S_2$ , note that (3.95) yields

$$P[f \cap C_e \neq \emptyset | N_e = x] \sim P[|f| \cap C_e = 1 | N_e = x] \sim \frac{d(x-d)}{n} \quad (d \leq x \leq N = o(n)). \quad (3.99)$$

Let us now estimate the probability that  $C_f$  is little given that  $|f| \cap C_e = 1$  and  $N_e = x \leq N$ . By Corollary 2 and Proposition 7, with probability  $\geq 1 - n^{-10}$  the random hypergraph  $H = H_d(n, p)$  has precisely one component  $\mathcal{C}$  of order  $\Omega_c(n)$ , while all other components have order  $O(\ln n)$ . Given that this is the case, the probability that  $f$  is little equals the probability that the set  $S = f \setminus C_e$  of cardinality  $s = d-1$  is little in the hypergraph  $H - C_e$ , which is distributed

Chapter 3. Counting Connected Graphs and Hypergraphs via the Probabilistic Method

as a random hypergraph  $H_d(n-x, p)$ . Since  $x \leq N = \tilde{O}(1)$ , we can apply Lemma 22 to the set  $S$  in  $H_d(n-x, p)$  to obtain that

$$\mathbb{P}[f \text{ is little} \mid |f| \cap C_e = 1, N_e = x] = (1-a)^{d-1} + n^{-\Omega(1)}. \quad (3.100)$$

By Lemma 22 and (3.91) we have  $|\theta - (1-a)^d| \leq n^{-\Omega(1)}$ . Hence, combining (3.99) and (3.100), we obtain

$$S_2 \sim \sum_{x=d}^N \left( (1-a)^{d-1} - (1-a)^d \right) \frac{d(x-d)}{n} \mathbb{P}[N_e = x] \sim a(1-a)^{d-1} d(\eta - d(1-a)^d) \cdot n^{-1}. \quad (3.101)$$

Finally, as  $\mathbb{P}[\text{both } e, f \text{ are little}] - \theta^2 \sim S_1 + S_2$ , the assertion follows from (3.98) and (3.101).

□

**Lemma 25** *Let  $e, f \subset V$ ,  $|e| = |f| = d$ ,  $|e \cap f| = 1$ . Then  $\mathbb{P}[\text{both } e, f \text{ are little}] \sim \zeta'$ , where  $\zeta' = (1-a)^{2d-1}$ .*

**Proof** Apply Lemma 22 to the set  $S = e \cup f$  of cardinality  $s = 2d - 1$ . □

**Corollary 5** *Let  $Z$  be the number of edges of  $H_d(n, p)$  that do not belong to the largest component. Then  $\text{Var}(Z) = O(m)$ .*

**Proof** Let  $e, f$  be variables that range over all subsets of  $V$  of cardinality  $d$ . Then (3.92) implies that

$$\begin{aligned} \text{Var}(Z) &= \sum_{e, f} \mathbb{P}[\text{both } e, f \text{ are little and } e, f \in E(H_d(n, p))] - \mathbb{P}[e \text{ is little, } e \in E(H_d(n, p))]^2 \\ &= \left[ p^2 \sum_{e \neq f} \mathbb{P}[\text{both } e, f \text{ are little}] - \mathbb{P}[e \text{ is little}] \right] + \mathbb{E}(Z) - \sum_e p^2 \mathbb{P}[e \text{ is little}]^2. \end{aligned}$$

Let  $S_j = \sum_{e,f:|e \cap f|=j} \mathbb{P}[\text{both } e, f \text{ are little}] - \mathbb{P}[e \text{ is little}] \mathbb{P}[f \text{ is little}]$ . Then Lemmas 24 and 25 entail that

$$S_0 p^2 = \binom{n}{d} \binom{n-d}{d} p^2 \zeta \sim \frac{n^2 c^2}{d^2} \zeta, \quad S_1 p^2 = \binom{n}{d} d \binom{n-d}{d-1} p^2 \zeta' \sim n c^2 \zeta'.$$

Moreover, for  $j \geq 2$  we have  $S_j p^2 = \tilde{O}(1)$ , and similarly  $\sum_e p^2 \mathbb{P}[e \text{ is little}]^2 = \tilde{O}(1)$ . Hence,  $\text{Var}(Z) \sim \mathbb{E}(Z) + \frac{n^2 c^2}{d^2} \zeta + n c^2 \zeta'$ . Thus, applying (3.3) to  $\zeta$  and  $\zeta'$  (cf. Lemmas 24 and 25), we obtain  $\text{Var}(Z) = O(m)$ .  $\square$

Since  $|E(H_d(n, p))|$  is binomially distributed with mean  $\binom{n}{d} p$ , the Chernoff bound (3.8) shows that  $|E(H_d(n, p))|$  is concentrated in width  $O(\sqrt{m})$  about  $\binom{n}{d} p$ . In addition, by Chebyshev's inequality and Corollary 5, the number  $Z$  of edges outside of the largest component is concentrated in width  $O(\sqrt{m})$  as well. Hence, also  $\mathcal{M}(H_d(n, p)) = |E(H_d(n, p))| - Z$  is concentrated in width  $O(\sqrt{m}) = O_c(\sqrt{bm})$ , so that (3.6) follows.

### 3.7.6 Proof of Lemma 14

To prove Lemma 14, we need the following technical statement.

**Lemma 26** *Suppose that  $1 \leq l = \tilde{O}(1)$ , and that  $0 \leq z \leq \tilde{O}(1)$ . Then*

$$1 - \tilde{O}(n^{-1}) \leq \binom{\binom{n-l}{d-1}}{z} p^z (1-p)^{\binom{n-1}{d-1}-z} \cdot [c^z \exp(-c)/z!]^{-1} \leq 1 + \tilde{O}(n^{-1}).$$

**Proof** Since  $\binom{\binom{n-l}{d-1}}{z} \leq \binom{n-1}{d-1}^z / z!$ , we have

$$\begin{aligned} \binom{\binom{n-l}{d-1}}{z} p^z (1-p)^{\binom{n-1}{d-1}-z} \cdot [c^z \exp(-c)/z!]^{-1} &\leq \left[ \binom{n-1}{d-1} p \cdot c^{-1} \right]^z \cdot \exp \left[ -\binom{n-1}{d-1} p + pz + c \right] \\ &\leq \exp(pz + c) = \exp(\tilde{O}(n^{-1})) = 1 + \tilde{O}(n^{-1}). \end{aligned}$$

Conversely, estimating  $\binom{\binom{n-l}{z}}{\binom{n-l}{d-1}} \geq (\binom{n-l}{d-1} - z)^z/z!$ , we obtain

$$\begin{aligned} & \left[ \binom{\binom{n-l}{z}}{\binom{n-l}{d-1}} p^z (1-p)^{\binom{n-l}{d-1}-z} \right]^{-1} \cdot c^z \exp(-c)/z! \\ & \leq \left[ c^{-1} \cdot \binom{n-1}{d-1} p \right]^{-z} \cdot \left[ \left[ \binom{n-l}{d-1} - z \right]^{-1} \cdot \binom{n-1}{d-1} \right]^z \exp \left[ \binom{n-1}{d-1} (p + O(p^2)) - c \right] \\ & \leq \left( \frac{(n-1)_{d-1}}{(n-l)_{d-1} - (d-1)!z} \right)^z \exp(O(pc)) \leq \left( \frac{n-1}{n - \tilde{O}(1)} \right)^{(d-1)z} \exp(\tilde{O}(n^{-1})) \leq 1 + \tilde{O}(n^{-1}), \end{aligned}$$

as claimed.  $\square$

To prove (3.65), let  $\mathcal{Z}_k$  signify the set of all sequences  $(z_1, \dots, z_k)$  of non-negative integers such that the sequence  $y_0 = 1, y_i = y_{i-1} + z_i - 1$  ( $i \geq 1$ ) satisfies  $y_i > 0$  for  $0 \leq i < k$  and  $y_k = 0$ . Then  $T = k$  iff  $(Z_1, \dots, Z_k) \in \mathcal{Z}_k$ , and similarly  $T' = k$  iff  $(Z'_1, \dots, Z'_k) \in \mathcal{Z}_k$ . Consequently, it suffices to prove that

$$\forall (z_1, \dots, z_k) \in \mathcal{Z}_k : \mathbb{P}[Z_i = z_i \text{ for } i = 1, \dots, k] = (1 + \tilde{O}(n^{-1})) \cdot \mathbb{P}[Z'_i = z_i \text{ for } i = 1, \dots, k]. \quad (3.102)$$

Indeed, in order to establish (3.102) we just need to prove that

$$\mathbb{P}[Z_i = z] = (1 + \tilde{O}(n^{-1})) \cdot \mathbb{P}[Z'_i = z] \quad \text{for } z = \tilde{O}(1). \quad (3.103)$$

For if (3.103) holds, then the fact that the random variables  $Z_i$  and  $Z'_i$  are mutually independent entails that for any sequence  $(z_1, \dots, z_k) \in \mathcal{Z}_k$  we have

$$\mathbb{P}[Z_i = z_i \text{ for } i = 1, \dots, k] = \prod_{i=1}^k \mathbb{P}[Z_i = z_i] = (1 + \tilde{O}(n^{-1}))^k \cdot \mathbb{P}[Z'_i = z_i \text{ for } i = 1, \dots, k],$$

which gives (3.102) if  $k = \tilde{O}(1)$ . Furthermore, as  $Z_i$  is distributed as  $\text{Po}(c)$  and  $Z'_i$  is distributed as  $\text{Bin}(\binom{n-l}{d-1}, p)$ , (3.103) follows immediately from Lemma 26, so that we have established (3.65). The proof of (3.66) is analogous.

## References

- [1] M. Ajtai and N. Linial. The influence of large coalitions. Technical Report 7133(67380),IBM, 1989.
- [2] N. Alon and M. Naor. Coin-flipping games immune against linear sized coalitions. In Proceedings of the IEEE Foundations of Computer Science(FOCS), 1990.
- [3] M. Ben-Or, N. Linial, and M. Saks. Collective coin flipping and other models of imperfect randomness. In Colloq. Math Soc. Janos Bolyai No., 52 Combinatorics, 1987.
- [4] M. Ben-Or and N. Linial. Collective coin flipping and other models of imperfect randomness. In Proceedings of the IEEE Foundations of Computer Science(FOCS), 1985.
- [5] M. Ben-Or and N. Linial. Collective coin flipping. Advances in Computing Research, pages 91-115, 1989. JAI Press; Silvio Micali, editor.
- [6] R. Boppana and B. Narayanan. Collective coin flipping and leader election with optimal immunity. Manuscript.
- [7] R. Boppana and B. Narayanan. The biased coin problem. In Proceedings of the Symposium on the Theory of Computing (STOC), 1993.
- [8] G. Bracha. An  $o(\log n)$  expected rounds randomized byzantine generals protocol. In Proceedings of the ACM Symposium on the Theory of Computation(STOC), 1985.
- [9] J. Cooper and N. Linial. Fast perfect information leader election protocol with linear immunity. *Combinatorica*, 15:319-332, 1995.
- [10] D. Dolev, M. Fischer, R. Fowler, N. Lynch, and H. Strong. An efficient algorithm for byzantine agreement without authentication. *Information and Control*, 1982.

## References

- [11] D. Dolev and R. Reischuk. Bounds on information exchange for byzantine agreement. In Proceedings of the First annual ACM symposium on Principles of distributed computing(PODC), 1982.
- [12] U. Feige. Noncryptographic selection protocols. In Proceedings of 40th IEEE Foundations of Computer Science(FOCS), 1999.
- [13] J. Kahn, G. Kalai, and N. Linial. The influence of random variables on boolean functions. In Proceedings of 29th IEEE Foundations of Computer Science(FOCS), 1988.
- [14] N. Linial. Games computers play: Game-theoretic aspects of computing. Technical report, Hebrew University of Jerusalem, 1992.
- [15] R. Ostrovsky, S. Rajagoplan, and U. Vazirani. Simple and efficient leader election in the full information model. In Proceedings of the Twenty Sixth Annual ACM Symposium on Theory of Computing, 1994.
- [16] A. Russell, M. Saks, and D. Zuckerman. Lower bounds for leader election and collective coin flipping in the perfect information model. In Proceedings of the Symposium on the Theory of Computing (STOC), 1999.
- [17] A. Russell and D. Zuckerman. Perfect information leader election in  $\log^*n + o(1)$  rounds. In Proceedings of 39th Annual Symposium on Foundations of Computer Science(FOCS), 1998.
- [18] M. Saks. A robust noncryptographic protocol for collective coin flipping. SIAM Journal of Discrete Mathematics, pages 240-244, 1989.
- [19] D. Zuckerman. Randomness-optimal oblivious sampling. Random Structures and Algorithms, 11:345-367, 1997.
- [20] V. Anand A. Bharathidasas. Sensor networks: An overview. Technical report, Dept. of Computer Science, University of California at Davis, 2002.
- [21] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. Wireless sensor networks: a survey. *Computer Networks*, 38:393–422, 2002.
- [22] Lichung Bao and J.J. Garcia-Luna. Transmission scheduling in ad hoc networks with directional antennas. In *ACM Int. Conference on Mobile Computing and Networking*, pages 48–58, Atlanta, 2002.
- [23] N. B. Clark, C. J. Colbourn, and D. G. Johnson. Unit disk graphs. *Discrete Mathematics*, 89:165–177, 1990.

## References

- [24] J. Díaz, J. Petit, and M. J. Serna. A random graph model for optical networks of sensors. *IEEE Transactions on Mobile Computing*, 2:143–154, 2003.
- [25] J. Díaz, V. Sanwalani, M. J. Serna, and P. Spirakis. Chromatic number of random scaled sector graphs. To appear in a special issue of Theoretical Computer Science devoted to the 2003 Dagstuhl Seminar on Graph Colorings.
- [26] J.M. Kahn, R.H. Katz, and K.S.J. Pister. Mobile networking for smart dust. In *ACM/IEEE Int. Conf. on Mobile Computing and Networking*, pages 176–189, Seattle, 1999.
- [27] C. McDiarmid. Random channel assignment in the plane. *Random Structures and Algorithms*, 22:187–212, 2003.
- [28] M Molloy and B. Reed. *Graph Coloring and the Probabilistic Method*. Springer, 2000.
- [29] M. Penrose. *Random Geometric Graphs*. Oxford Studies in Probability, Oxford U.P., 2003.
- [30] N. Alon and J. Spencer. The probabilistic method. 2nd edition. Wiley 2000.
- [31] D. Barraez, S. Boucheron, and W. Fernandez de la Vega. On the fluctuations of the giant component. *Combinatorics, Probability and Computing* **9** (2000) 287–304.
- [32] E.A. Bender, E.R. Canfield, and B.D. McKay. The asymptotic number of labeled connected graphs with a given number of vertices and edges. *Random Structures and Algorithms* **1** (1990) 127–169.
- [33] E.A. Bender, E.R. Canfield, and B.D. McKay. Asymptotic properties of labeled connected graphs. *Random Structures and Algorithms* **3** (1992) 183–202.
- [34] B. Bollobás. *Random graphs*, 2nd edition, Cambridge University Press 2001.
- [35] A. Coja-Oghlan, C. Moore, and V. Sanwalani. Counting Connected Graphs and Hypergraphs via the Probabilistic Method. APPROX-RANDOM 2004: pages 322-333.
- [36] W. Feller. *Introduction to probability theory and its applications*. Wiley 1968.
- [37] R. van der Hofstad and J. Spencer. Counting connected graphs asymptotically. Preprint (2005).
- [38] S. Janson. The minimal spanning tree in a complete graph and a functional limit theorem for trees in a random graph. *Random Structures and Algorithms* **7** (1995) 337–356.

## References

- [39] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*, Wiley 2000.
- [40] M. Karoński and T. Łuczak. The phase transition in a random hypergraph. *J. Comput. Appl. Math.* **142** (2002) 125–135.
- [41] M. Karoński and T. Łuczak. The number of connected sparsely edged uniform hypergraphs. *Discrete Math.* **171** (1997) 153–168.
- [42] T. Łuczak. On the number of sparse connected graphs. *Random Structures and Algorithms* **1** (1990) 171–173.
- [43] N. O’Connell. Some large deviation results for sparse random graphs. *Prob. Th. Relat. Fields* **110** (1998) 277–285.
- [44] B. Pittel. On tree census and the giant component in sparse random graphs. *Random Structures and Algorithms* **1** (1990) 311–342.
- [45] B. Pittel and N.C. Wormald. Counting connected graphs inside out. to appear in *J. Combinatorial Theory Series B*.
- [46] J. Schmidt-Pruzan and E. Shamir. Component structure in the evolution of random hypergraphs. *Combinatorica* **5** (1985) 81–94.